

NBER WORKING PAPER SERIES

MARKET DESIGN IN CAP AND TRADE PROGRAMS:
PERMIT VALIDITY AND COMPLIANCE TIMING

Stephen P. Holland
Michael R. Moore

Working Paper 18098
<http://www.nber.org/papers/w18098>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2012

The authors thank Severin Borenstein, Garth Heutel, Richard Newell, Billy Pizer, Stephen Salant, and Andrew Yates and seminar participants at UNCG and Camp EI@Haas. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2012 by Stephen P. Holland and Michael R. Moore. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Market Design in Cap and Trade Programs: Permit Validity and Compliance Timing
Stephen P. Holland and Michael R. Moore
NBER Working Paper No. 18098
May 2012
JEL No. H4,Q4,Q5

ABSTRACT

Cap and trade programs have considerable heterogeneity in permit validity and compliance timing. For example, permits have different validity across time (e.g., banking, borrowing, and seasons) and space (e.g., zonal restrictions), and compliance timing can be annual, in overlapping cycles, or in multi-year periods. We compare and contrast nine prominent cap and trade programs along these dimensions and construct a general model of permit validity and compliance timing. We derive sufficient conditions under which abatement is invariant to compliance timing, i.e., compliance timing cannot smooth abatement cost shocks. Under these conditions, i) expected compliance costs are invariant, ii) the variance of compliance costs increases with delayed compliance, iii) equilibrium prices may not be unique, and iv) the delayed compliance equilibrium may rely upon non-unique, “degenerate” prices not determined by marginal abatement costs. Degenerate prices are unlikely to be discovered by market forces. We then present two examples which are not invariant to compliance timing. If permit allocation is delayed or if a price cap is implemented with a reserve fund, abatement may depend on compliance timing. We demonstrate the model’s broad applicability by illustrating different types of temporal and spatial permit validity.

Stephen P. Holland
Bryan School of Business and Economics
University of North Carolina, Greensboro
P.O. Box 26165
Greensboro, NC 27402-6165
and NBER
sphollan@uncg.edu

Michael R. Moore
School of Natural Resources & Environment
University of Michigan
Ann Arbor, MI 48109-1041
micmoore@umich.edu

1. Introduction

As the number of cap and trade programs for reducing pollution has increased, so has their complexity along a number of dimensions. Programs now have a variety of rules regarding emissions permit validity across time and space. For example, some permits can be banked for future use; some can be borrowed for current use; some have overlapping applicability periods; and some are restricted to certain regions within the overall market. Similarly, a variety of rules exist for compliance timing and “true-up” of emissions obligations.¹ Some programs require annual compliance, some have overlapping compliance cycles, and some delay compliance as much as four (!) years. Given this breadth of approaches to permit validity and compliance timing, it is imperative to understand *i)* which aspects of permit validity and compliance timing affect market outcomes and compliance costs; and *ii)* the precise mechanisms for these effects. This paper explores these questions by analyzing a general model of permit validity and compliance timing.

Flexibility in permit validity across time and regions is a valuable tool for reducing abatement costs.² Whenever sources have different marginal abatement costs, a cost reduction opportunity exists. Making permits valid for a wider range of compliance obligations allows the program to realize more of these cost reduction opportunities by trading. However, the regulator may wish to prohibit some trades if the marginal damages are different across the two sources.³ Since requiring approval of individual trades increases transaction costs, regulators generally define broad classes of permits (vintages) with identical applicability criteria. Permit vintages allow emissions trading programs to realize many cost reduction opportunities while avoiding pollution “hot spots”.⁴

¹ We use “compliance” and “true up” to refer to the process of reconciling a regulated facility’s observed emissions with its permit holdings.

² The literature on trading across time (banking and borrowing) is vast, e.g., Kling and Rubin (1997), Schennach (2000), and Yates and Cronshaw (2001). See Montgomery (1972) and Fowlie and Muller (2010) for analysis of trading across regions.

³ Specifically, the regulator would prohibit a trade if the difference in marginal damages between the source increasing emissions and the source decreasing emissions is greater than the difference in marginal abatement costs.

⁴ Muller and Mendelsohn (2009) argue that marginal damages should be incorporated directly into the emissions trading program.

Similarly, flexibility in compliance timing and in true-up procedures is generally thought to increase liquidity, to reduce compliance costs, and to provide flexibility to adjust to cost shocks.⁵ Several programs have been quite creative in introducing flexibility into compliance timing and true-up. For example, one program introduced two separate compliance cycles based on the fiscal year and the calendar year. Another program simply delays compliance until three or four years after the start of the program. However, the precise mechanism by which delayed compliance could affect compliance costs has not been specified.

Section 2 describes compliance timing; temporal and spatial restrictions on permit validity; and permits-to-emissions compliance ratios across nine major cap and trade programs. The programs have considerable heterogeneity across these dimensions. There is dramatic variation in compliance timing, ranging from i) every year to ii) once every three or four years to iii) overlapping compliance cycles to iv) partial compliance requirements. Temporal restrictions on permit validity include provisions allowing (or disallowing) i) banking, ii) borrowing, iii) borrowing with interest, and iv) borrowing within a compliance period. Spatial restrictions on permit validity include implicit restrictions through the limited geographic scope of a program and explicit limits on trading within the program. Finally, programs have begun to experiment with different permits-to-emissions compliance ratios to impose both temporal and spatial restrictions on pollution. The heterogeneity across each of these dimensions is described in Section 2.

Section 3 presents a general model of permit validity and compliance timing that allows for considerable flexibility along these dimensions. The first result from the model derives sufficient conditions for the equilibrium to be invariant to compliance timing. The first sufficient condition is that the current price equals the present value of the expected future price. This is a standard arbitrage condition. The second sufficient condition is quite general and applies to a wide range of program types. For example, it applies if vintages of permits can be ranked by their applicability, such as with banking, so that some permits are always more valuable than other.

⁵ See the discussions of compliance timing in RECLAIM and RGGI in Section 2.

The intuition of invariance to compliance timing follows from the information contained in the current price. With prompt compliance, abatement is based on the current price. With delayed compliance, abatement is based on the present value of the expectation of the price at compliance time. If arbitrage equates the current price to the discounted expected future price, then abatement is invariant to compliance timing.

Although abatement is invariant to compliance timing, compliance costs are not invariant. The second result shows that, under the sufficient conditions, expected compliance costs are invariant to delayed compliance, but the variance of compliance costs increases with delayed compliance. Expected compliance costs are invariant since the prompt compliance price equals the discounted expected price of delayed compliance. However, with delayed compliance, the compliance costs may be either higher or lower than the expected cost. Thus the variance increases.

Analysis of the model also shows that the equilibrium can have non-unique and “degenerate prices”: equilibrium prices which are not determined by sloping supply and demand functions, i.e., prices which are unlikely to be discovered by market forces. The third result shows that the non-unique and degenerate prices are relied upon in transactions under delayed compliance, but are irrelevant under prompt compliance. The problem arises since the only elasticity in emissions markets occurs at the time the emissions take place. Once the emissions leave the source, the emissions (and resulting compliance obligation) are sunk. Since demand and supply are each perfectly inelastic at a delayed compliance date, a variety of prices could clear the market. The equilibrium prices can be determined by future abatement cost shocks and an arbitrage condition. Degenerate prices are determined by the arbitrage condition (rather than by abatement costs) and are prices that justify *ex post* the *ex ante* price expectation.

Despite the generality of the compliance invariance results, the sufficient conditions do not hold for all possible programs. Section 3 presents two examples where invariance fails. In the first example, the allocation of permits is delayed, and this delayed allocation can lead to abatement which depends on compliance timing. The intuition follows from a simple two year market. If very few permits are allocated in the first year and compliance is prompt, then the first period permit price would be very high. Delaying compliance would allow the permits allocated in the second year to be used for

the first period emissions and would lower the first period price. In this example, delayed compliance reduces abatement costs, but does not reduce them below the level that would have resulted with prompt allocations.

The second example involves a price ceiling supported by a reserve fund. The analysis, which is based closely on Hasegawa and Salant (2010a and 2010b), shows that the equilibrium may depend on compliance timing. To illustrate the intuition, suppose all the permits are in the reserve fund and the reserve fund is large. Since the reserve fund is large, the price will not exceed the ceiling. With delayed compliance, abatement is based on the present value of the price ceiling, which is lower than the price ceiling. However, with prompt compliance, abatement is based on the current price which must equal the price ceiling. Thus the price is higher and emissions are lower with prompt compliance, i.e., abatement depends on compliance timing.

The generality of the model is illustrated in Section 4 using several specific features of temporal and spatial permit validity, including: banking, borrowing, borrowing with interest, spatial segmentation, and permits-to-emissions compliance ratios. Intertemporal prices grow at the rate of interest, following the arbitrage condition, in the four temporal illustrations. Features of spatial validity can be used to decentralize the efficient permit allocation. The set of illustrations demonstrate the model's broad applicability to market designs in practice.

The compliance invariance results show that delayed compliance cannot provide flexibility for responding to cost shocks, since the equilibrium, even with cost shocks, is invariant to whether compliance is prompt or delayed. This result undercuts one of the primary rationales for delayed compliance, and suggests that regulators need to rely upon broader permit validity and/or hybrid price containment mechanisms (e.g., reserve funds) to provide flexibility.

The compliance invariance results are perhaps reassuring in light of the variety of approaches programs take toward compliance timing. However, increased variance of compliance costs, the reliance on non-unique and degenerate prices, and the potential interactions with a reserve fund are factors that regulators would analyze carefully when considering delayed compliance. Regulators can also assess other factors such as administrative costs, salience of compliance costs, and complications from bankruptcy

when considering compliance timing. These conclusions are described more fully in Section 5.

2. Provisions for Compliance Timing and Permit Validity in Cap and Trade Programs

Cap and trade programs contain considerable differences in provisions for compliance timing and permit validity across time and space. This section compares and contrasts compliance timing, permit validity, and permits-to-emissions compliance ratios in nine major cap and trade programs: the U.S. SO₂ market (Acid Rain Program, Title IV of Clean Air Act, or ARP)⁶; the U.S. NO_x market (NO_x Budget Trading Program or NBP)⁷; the Clean Air Interstate Rule (CAIR) and the Cross-State Air Pollution Rule (CSAPR)⁸, both of which regulate NO_x and SO₂ in the eastern U.S.; the Southern California NO_x market (RECLAIM)⁹; the European Union's CO₂ market (Emissions Trading System, or EU ETS)¹⁰; a legislative proposal in the U.S. Congress to establish a U.S. CO₂ market (Waxman-Markey H.R. 2454)¹¹; the northeastern U.S. CO₂ market (RGGI)¹²; and the California CO₂ market (AB 32). Table 1 outlines these provisions for each program.

2.1 Compliance Timing

The most common approach to compliance timing is for true up on an annual basis. The ARP, NBP, CAIR, EU ETS, and CSAPR all have annual compliance, and Waxman-Markey would likely have required annual compliance. With annual compliance, monitoring and reporting generally occur throughout the year.¹³ Regulated facilities are required to operate monitoring equipment, such as a continuous emissions

⁶ See Ellerman et al. (2000), Joskow et al. (1998), Stavins (1998), and USEPA (2009a and 2009b).

⁷ See USEPA (2005a, 2005b, and 2009c).

⁸ See Federal Register (2005) and USEPA (2011).

⁹ RECLAIM is the acronym for the Regional Clean Air Incentives Market, which operates in the greater Los Angeles metropolitan area. See Holland and Moore (2012), SCAQMD (2009), and USEPA (2006).

¹⁰ See Ellerman and Joskow (2008), European Commission (2003), Kruger and Pizer (2004)

¹¹ Waxman-Markey passed the House. Several similar proposals in the Senate, including Kerry-Boxer and McCain-Lieberman, were never voted on. See U.S. House of Representatives (2009).

¹² RGGI is the acronym for the Regional Greenhouse Gas Initiative, which is an agreement among ten states in the northeastern and mid-Atlantic region. See RGGI (2007 and 2008).

¹³ For the ARP, monitoring and reporting requirements, known as "Part 75" are detailed in 40 CFR 75. The same requirements have been adopted for other programs such as CAIR.

monitoring system (CEMS), which records hourly emissions and/or heat input. Each facility then submits a regular report of its emissions to the regulator who verifies the emissions report and reconciles any discrepancies. For the ARP, the quarterly emissions report is submitted electronically, and the first verification screen occurs electronically at the time of filing. Any discrepancies in the report should be resolved within 30 days of the deadline for filing the quarterly report. After the end of the year, there is generally a trading period in which regulated facilities can buy or sell permits. At the end of the trading period (for example, by March 1 for the ARP¹⁴), each regulated facility must submit an annual compliance report verifying that all monitoring and reporting requirements have been satisfied and specifying which permits are to be used for compliance.¹⁵ The facility's account must have sufficient permits to cover the preceding year's emissions, and these permits are later deducted from the account. This process is sometimes referred to as "true-up". If there are insufficient permits in the account to cover the obligation, the facility is considered to be out of compliance and penalties are assessed. Generally, compliance rates are quite high.¹⁶

Three programs—RGGI, RECLAIM, and AB 32—have approaches to compliance timing which are substantially different. RECLAIM has two overlapping compliance cycles, each of which covers a 12-month period (SCAQMD 2009), with roughly equal numbers of facilities in each cycle. Cycle 1 is based on the calendar-year (Jan. 1 - Dec. 31), and Cycle 2 is based on the fiscal-year (July 1 - June 30).¹⁷ Emissions are recorded hourly, daily, or monthly depending on the size of the facility and are reported quarterly by all facilities. For both compliance cycles, there is a 60-day reconciliation period at the end of the respective cycle. Thus compliance and true-up occurs for half the facilities in August and for half the facilities in February.

RECLAIM adopted overlapping compliance cycles to reduce administrative costs; to increase liquidity and flexibility; and to allow firms to respond better to abatement cost

¹⁴ 40 CFR Part 72.

¹⁵ If different vintage permits could be used to cover the obligation, the regulator either uses some rule for determining which permits to deduct or allows the facility to specify which permits to deduct.

¹⁶ Schakenbach et al. (2006) claim compliance above 99% in ARP and NBP.

¹⁷ Emissions permits are similarly defined (i.e., have similar overlapping validity dates) but can be used by facilities in either compliance cycle.

shocks.¹⁸ The overlapping compliance cycles mirror the overlapping permit cycles (discussed below) and these two design features are frequently confused.¹⁹ Carlson et al. (1993) carefully analyze both staggered compliance dates and staggered issue dates, but do not experimentally analyze staggered issue dates without staggered compliance dates. They find experimental evidence that staggered issue dates and compliance dates prevent price spikes, and recommend both staggered issue dates and staggered compliance dates.

In the early years of the program (1994-1996), there were several cases where facilities did not hold sufficient permits in their accounts at the required dates and later were assessed monetary and permit penalties. These early instances of non-compliance were likely exacerbated by administrative procedures on the part of RECLAIM (USEPA 2002). The emissions market was not substantially affected by these early violations since the market had an excess supply of permits. Since 1997, there has been much better compliance (rates typically in the 94-97% range) with no substantial violations.²⁰

Compliance in RGGI is based on control periods. The first control period begins January 1, 2009 (RGGI 2007, RGGI 2008). The control period is scheduled to be three years but can be extended to a fourth year if a trigger event is declared (i.e., if the twelve-month rolling average CO₂ allowance price exceeds \$10 adjusted for inflation.) The monitoring and reporting requirements are based on those pioneered by ARP: i.e., each facility is required to have monitoring equipment and to submit quarterly emissions reports within 30 days of the end of the quarter. After the control period ends, there is a transfer deadline on March 1 of the following year. By this date, sufficient permits must be in the account in order for the facility to be in compliance for the preceding control period. Note that under this system, a facility may not need to purchase the necessary permits until up to four years (!) after the emissions occurred.

¹⁸ According to Carlson and Scholtz (1994), “staggered compliance periods smooth facilities’ reactions to unexpected events, resulting in pollution periods that should be less chaotic and less likely to yield NAAQS violations; businesses will find the adjustment to tighter standards easier. Ultimately, the risks inherent in pollution management are reduced.”

¹⁹ For example, Carlson and Scholtz (1994) implicitly recognize the difference between permit validity and compliance timing when they note that facilities could receive a “mixed” allocation of permits. However, their arguments for overlapping compliance cycles mix rationales for overlapping compliance cycles and overlapping permit cycles.

²⁰ Compliance rates are reported in the various issues of the RECLAIM Annual Report (http://www.aqmd.gov/reclaim/reclaim_annurpt.htm).

The multi-year compliance periods were adopted to provide flexibility for adjusting to cost variations, to reduce administrative costs, and to allow “de facto” borrowing.²¹ Extended compliance periods are serving as a model for other programs, e.g., the proposed Western Climate Initiative (2009) and Midwest Greenhouse Gas Reduction Accord (2009) define three-year compliance periods in their draft model rules.

AB 32 has two forms of compliance timing: annual and triennial (California Environmental Protection Agency, 2010). The first three-year compliance period is slated for 2012-2014. For triennial compliance, facilities must have enough permits in their accounts by November 1 of the following year to cover all their emissions over the preceding three years. Any permit which is valid at that date can be used to cover any obligation for triennial compliance. Annual compliance essentially requires a “down payment” on true-up. By July 15 of each year, each facility must have sufficient permits in its account to cover thirty percent of their emissions obligation from the preceding year. For this down payment, the allocation year of the permit must be the current year or an earlier year. For example, a 2013 permit is *not* valid for covering the annual down payment on 2012 emissions but would be acceptable for covering the remaining balance beyond the down payment at triennial compliance. Annual emission reports are required, and AB 32 includes a novel element of third-party verification of each report.

2.2 Permit Validity

In the simplest cap and trade program, any permit could be used by any facility at any time to cover any emissions obligation, and all permits would be perfect substitutes. No program is this simple, since unrestricted trading could allow temporal or spatial hotspots in pollution. Every program limits the validity of permits in some way, such that not all permits are perfect substitutes. To facilitate trading, programs generally create groups of permits which can be used to satisfy the same emissions obligations and are perfect substitutes. We call these groups of permits *vintages*.

²¹ “Multi-year compliance periods were employed to provide regulated facilities more flexibility to adjust to variations in electricity demand (driven by meteorology and load growth), fuel price spikes, clean unit outages, etc. A longer compliance period may also lead to resource (administrative) savings for the regulated facilities and the states implementing the program. This design component was included in lieu of allowance borrowing, as it allows for de facto borrowing within a three-year compliance period.” RGGI (2007)

The four key features of a permit vintage are the permit's expiration date (the last date of emissions for which the permit can be used), its start date (the first date of emissions for which the permit can be used), its spatial applicability (which facilities can use it), and how much emissions each permit covers. In practice, permits are generally labeled with an *allocation year*, and expiration and start dates are determined by whether a permit can be banked (i.e., used after its allocation year) or borrowed (i.e., used before its allocation year). The programs take a variety of approaches to temporal restrictions (through banking and borrowing), spatial limitations, and permits-to-emissions compliance ratios.

In program design, permit validity involves a trade-off between abatement costs and damages. Broader permit validity allows for more trading which reduces abatement costs. However, broader permit validity can allow pollution "hot-spots" either temporally or spatially. Thus programs to control pollutants with more temporal and spatial heterogeneity in damages (e.g., ozone) would have more restrictions on permit validity, whereas permits for pollutants with less temporal and spatial heterogeneity (e.g., CO₂) would have fewer restrictions on permit validity.

Temporal restrictions in permit validity

Banking provisions are generally much less restrictive than borrowing provisions. For example, ARP, CAIR, and CSAPR have unlimited banking but no borrowing. One reason for this may be that borrowing can affect a program's credibility. For example, if borrowing were excessive, the limited supply of remaining permits (and resulting price spike) would likely force regulators to intervene.²² However, the likelihood of intervention can make excessive borrowing optimal. In the extreme case of unlimited borrowing, the equilibrium could have a permit price of zero (i.e., no abatement) followed by a forced regulator intervention. Unlimited banking does not raise similar credibility issues.

²² For example, RECLAIM administrators were forced to intervene in 2001 when the permit price spiked, see Holland and Moore (2012).

Despite concerns about program credibility, some programs have attempted to incorporate borrowing since borrowing can decrease abatement costs. In particular, the proposals regulating CO₂ in the US Congress allow some restricted borrowing. For example, Waxman-Markey, which has unlimited banking, allows borrowing only from allocations up to five years in the future and only for up to 15 percent of a regulated facility's annual emissions. The program further limits borrowing by requiring the regulated facility to retire additional permits as "interest" payments. Borrowing from the next year's allocation occurs without interest. However, borrowing from vintages two to five years in the future requires an interest payment computed by multiplying 0.08 times "the number of years between the calendar year in which the allowance is being used ... and the vintage year of the allowance".^{23, 24}

RGGI, which also regulates CO₂ and has unrestricted banking, only allows borrowing within the three-year control period. According to a program overview, "This design component was included in lieu of allowance [permit] borrowing, as it allows for de facto borrowing within a three-year compliance period" (RGGI 2007). Thus, permits can only be borrowed from at most four years in the future in the event of a declared emergency.

California's AB 32 has similar provisions to RGGI: unlimited banking along with borrowing of annual allowances within a three-year compliance period. AB 32's annual requirement for a thirty percent down payment on emissions compliance limits borrowing in each year to no more than seventy percent of the emissions obligation. This generous limit on borrowing is unlikely to be binding.

The EU-ETS, which also regulates CO₂, placed restrictions on both banking and borrowing. Permits from its first period (2005-07) could be banked but only until the end of this period. Permits from the second period (2008-12) have unlimited banking. Borrowing is allowed but only implicitly through a compliance provision. A permit can be borrowed forward by one year, as "allowances [permits] for each year are to be issued

²³ For example, if the calendar year is 2012 and the borrowed vintage year is 2016, the interest payment is 0.32 [=0.08*(2016-2012)] permits per borrowed permit.

²⁴ These interest payments are formally equivalent to an increased permit-to-emissions compliance ratio as discussed below.

before February 28, while compliance for the previous year is assessed after April 30. This allows de facto borrowing between January and May, as participants can use this period to make up for a deficit in their holdings” (Convery and Redmond 2007).²⁵

This market design led to some interesting price dynamics in the EU-ETS. Despite the fact that permits from the two periods are not substitutable in any way, the prices tracked quite closely until 2007. On April 25, 2006, the price of the first-period vintages crashed eventually falling to virtually zero while second-period vintages traded in the €15-25 range. (Ellerman and Joskow 2008; Bushnell, Chong, and Mansur 2009).

Markets for NO_x (an ozone precursor) must be more careful about temporal hotspots and tend to have more restrictions on both banking and borrowing. The NBP does not allow any borrowing and restricts banking through a “flow control” provision (EPA 2005a, EPA 2005b). Under this provision, banked permits cover emissions at a 1:1 ratio until the total number of banked permits exceeds 10% of the yearly permit allocation. If this flow control provision is triggered, a flow control ratio is calculated.²⁶ The flow control ratio is then applied to each facility’s account of banked permits at compliance time and determines the proportion of the banked permits which cover emissions at a 1:1 ratio and at a 2:1 ratio.²⁷ Thus flow control reduces the value of banked permits when the trigger is exceeded. The flow control trigger has regularly been exceeded and banked permits have been used at the 2:1 ratio.²⁸ CAIR did not continue the NBP’s flow control provision after 2009, and neither will CSAPR beginning 2012.

Another NO_x market (RECLAIM) ruled out banking “because of concerns that the ability to use banked emissions might lead to substantial increases in actual emissions in some future year, and thus delay compliance with ambient air quality standards”

²⁵Ellerman and Trotignon (2009) develop evidence of both borrowing and banking in the EU ETS during 2005-2007.

²⁶ The flow control ratio is defined as

$$\frac{10\% * \text{permit allocation for year } t}{\text{aggregate banked permits at compliance time } t-1}$$

Note that the flow control ratio is between 0 and 1, and equals 1 if the flow control trigger is just binding.

²⁷ For example, suppose the flow control provision is triggered and the flow control ratio is 0.25. If a facility has 1000 banked permits, these banked permits could cover only 625 tons of emissions (250+750/2).

²⁸ See <http://www.epa.gov/airmarket/progress/progress-reports.html>.

(Ellerman, Joskow, and Harrison 2003, 21). Similarly, borrowing was not allowed. However, the program designers did attempt to allow some limited intertemporal trading through overlapping permit cycles. The overlapping permit cycles mimic RECLAIM's overlapping compliance cycles described above. In particular, all permits are valid for twelve months, but some permits expire in December and other expire in June. Importantly any facility can use any unexpired permit, although facilities were only allocated permits whose validity dates match their compliance cycle. Holland and Moore (2012) model the scope for intertemporal trading in RECLAIM and show empirical evidence consistent with intertemporal trading.

Banking and borrowing create explicit limitations on intertemporal trading. However, some programs create implicit limitations on intertemporal trading through the scope of the program. Most prominently, NBP was only in effect during the summer ozone season. The limited scope of the NBP implicitly prevented trading between the summer ozone season and the winter. CAIR and CSAPR, which were built on NBP, then established two distinct seasonal NO_x markets. The summer ozone season program regulates NO_x emissions during the summer when marginal damages from ozone are highest. The annual program regulates NO_x emissions year-round to manage particulate formation. These overlapping regulatory seasons continue this barrier to intertemporal NO_x trading.

Spatial Restrictions in Permit Validity

Some programs, especially those regulating local pollutants, place restrictions on the geographic validity of permits. Since CO₂ is a global pollutant, none of the programs regulating CO₂ define explicit spatial restrictions. Other programs limit spatial trading either implicitly by the geographic scope of the program (e.g., RECLAIM applies only to the Los Angeles region and CAIR only applies to the eastern U.S.) or explicitly by restricting permit validity.²⁹

In addition to the narrow geographic scope of program, RECLAIM defines two

²⁹ Fowlie and Muller (2010) calculate the loss from not incorporating spatial variations in damages in NBP. They estimate that net benefits could increase by 17% by incorporating more spatial heterogeneity in the program.

spatial zones in the greater Los Angeles region: coastal and inland. Due to the natural airflow from west to east, trading permits from the inland to the coastal region would increase coastal pollution but may not reduce inland pollution. Conversely, trading permits from the coastal to the inland region would increase inland pollution but would definitely decrease coastal pollution. To discourage trading permits from the inland to the coastal region, regulators assigned facilities to coastal and inland zones and similarly designated permits as coastal or inland. Inland facilities can use either permits, whereas coastal facilities can only use coastal permits. This explicit spatial limitation reduces the local damages of trading.

CAIR supplements ARP in regulating SO₂ emissions by reducing the geographic scope of the program. While ARP covers the conterminous United States, CAIR covers a smaller region of 28 eastern states and the District of Columbia. This narrower geographic scope—together with higher permits-to-emissions compliance ratios, discussed below—is meant to address the regional contributions of SO₂ emissions to local particulate matter pollution.

Under CAIR, the two temporally segmented NO_x programs are also segmented spatially. Twenty-two eastern states operate with both the seasonal and the annual NO_x programs. An additional three states are in the seasonal program, and another three states are in the annual program. As mentioned above, the temporal and spatial segmentation of the NO_x programs is due to different air quality issues: ozone and particulate matter.

As the replacement of CAIR, CSAPR continues a similar spatial structure for the two NO_x programs (USEPA, 2011). It also creates two spatially distinct SO₂ markets in the eastern United States. One market covers 16 states (“Group 1” states) with facilities subject to relatively higher emission reductions. A second market covers 7 states (“Group 2” states) with facilities subject to relatively lower emission reductions. Permits allocated or defined for one market cannot be used for compliance with the other market.

To limit leakage in the electricity sector, CO₂ emissions from the generation of electricity imported to California are regulated under AB 32 (California Environmental Protection Agency, 2010). A firm that holds title to the electricity as it enters the state is

the regulated entity. This provision is unusual in that it enlarges rather than restricts the geographic validity of permits.

2.3 Permits-to-Emissions Compliance Ratios

In cap and trade programs, facilities do not trade “emissions” but rather permits (e.g., “title IV SO₂ allowances” and “RECLAIM Trading Credits (RTCs)”). Typically one permit offsets one ton of pollution: e.g., one title IV SO₂ allowance is retired for each ton of SO₂ emissions and one RTC is retired for each ton of NO_x emissions. However, some programs have modified this simple 1:1 linkage in order to introduce more flexible spatial and temporal limitations. We call these restrictions *permits-to-emissions compliance ratios*. Most programs (e.g., ARP, RECLAIM, RGGI, and the EU-ETS) have 1:1 compliance ratios; however some programs have higher compliance ratios.

One example is the NBP. As discussed above, if the flow control provision is triggered a proportion of the facility’s banked permits can only be used at a 2:1 ratio rather than at a 1:1 ratio. Another example is “borrowing with interest” in Waxman-Markey, which is essentially an increased compliance ratio.

CAIR uses compliance ratios to reduce SO₂ emissions in the eastern United States. The CAIR SO₂ program, which was built on ARP and used ARP permits, requires higher compliance ratios for permits with vintages after 2010. In particular, permit allocations for 2010 to 2014 have a 2:1 compliance ratio and permit allocations beyond 2014 have a 2.86:1 compliance ratio. This provision effectively reduces the cap by increasing the compliance ratio above 1:1.³⁰ SO₂ permits continue to cover emissions at the 1:1 ratio specified in ARP in states outside of the CAIR region.

CSAPR also introduces the possibility of a higher compliance ratio through an “assurance provision” on SO₂ and NO_x emissions at the state level (USEPA, 2011). A “variability limit” is set at a state’s annual allocation of permits plus an additional 18 percent of the allocation for the annual program and plus 21 percent for the seasonal NO_x program. If emissions from facilities in the state exceed the variability limit, a 2:1

³⁰ Interestingly, the effect on permit prices is ambiguous: each permit is less valuable, but equilibrium marginal abatement cost increases.

compliance ratio switches on; offending facilities must surrender two permits, not one, for each ton of excess emissions. This assurance provision is intended to remedy a legal weakness of CAIR by providing disincentive for interstate trading.

3. A Model of Permit Validity and Compliance Timing

To analyze permit validity and compliance timing, we develop a model of a perfectly competitive emissions permits market. After deriving the main results on compliance timing invariance, we also consider cases in which equilibria can vary with compliance timing.

3.1 The model

Consider a model of permit validity and compliance timing in a cap and trade program. The regulator creates and defines emissions permits which participating facilities use to satisfy their compliance obligations.³¹ Define a *vintage* as a class of permits with identical applicability criteria.³² Since permits within a vintage are perfect substitutes, they trade at a common price. Let p_{jt} be the price of a vintage j permit in period t .³³ Assume the regulator issues \bar{E}_j permits of vintage j .

Let e_{it} be emissions from facility i in period t , and let the abatement cost function be $c_{it}(e_{it}; \theta_{it})$ where θ_{it} is a stochastic abatement cost shock.³⁴ In each period, the facility realizes its abatement cost shock and decides how much to abate. Assume that abatement costs are convex in emissions with $c_{it} > 0$; $c'_{it} < 0$; and $c''_{it} > 0$, i.e., marginal abatement costs, $-c'_{it}$, are positive and increasing in abatement, decreasing in emissions.³⁵

³¹ The permits can be grandfathered or auctioned. For now, we assume permits are allocated at the start of the program.

³² To be precise, if one permit in vintage j can be used by facility i to satisfy a given compliance obligation from emissions at time t , then any permit in vintage j can be used to satisfy the same compliance obligation.

³³ The period of analysis can be quite general, e.g., monthly or yearly, but we assume the first period of the program is $t=1$.

³⁴ Appendix A presents the simpler model without the abatement cost shock.

³⁵ Emissions control technology may require capital investments. Delaying these investments may have an option value. Here we focus on marginal (rather than fixed) abatement costs which determine the prices of emissions permits.

To model compliance timing, let $\phi_i(t)$ be facility i 's compliance time for period t emissions where $\phi_i(t) \geq t$, i.e., compliance cannot take place before emissions occur. We distinguish *prompt compliance* from *delayed compliance*, with prompt meaning that compliance is required in each period and delayed pertaining to all other cases. By definition, $\phi_i(t) = t$ with prompt compliance. In a program where compliance first occurs three years after the start of the program, i.e., similar to RGGI, then $\phi_i(1) = 3$; $\phi_i(2) = 3$; and $\phi_i(3) = 3$, and the program requires delayed compliance. Note that the compliance time need not be the same for all facilities (e.g., RECLAIM has two separate compliance cycles).

In a compliance period, the facility must determine how many permits of each vintage to utilize to cover past emissions. Let $\phi^{-1}(t)$ be the set of periods for which compliance occurs in t , i.e., $\phi^{-1}(t) \equiv \{t' | \phi_i(t') = t\}$. Note that ϕ_i^{-1} is a correspondence and is the “inverse correspondence” of the function ϕ_i . Further note that $\phi_i^{-1}(t)$ can be the empty set, in which case t is not a compliance period. If $t' \in \phi^{-1}(t)$, let $d_{ijt'}$ be facility i 's demand for permit vintage j to cover period t' emissions. This demand need not be determined until the compliance period t .

To model permit validity, define the *compliance factor* as the amount of emissions obligations covered by one permit. Specifically, let the compliance factor α_{ijt} be the amount of facility i 's period t emissions covered by one permit of vintage j . The factor α_{ijt} is very general and can model spatial limits, temporal limits, different permits-to-emissions compliance ratios, and negative marginal damages (see Fowlie and Mueller 2010).³⁶ Importantly, if facility i cannot use permit vintage j for emissions at time t due to spatial or temporal limits, then $\alpha_{ijt} = 0$.

Each facility minimizes compliance costs by choosing emissions in each period and how many permits of each vintage to use. Let facility i 's value function in period t be $V_{it}(\bar{e}_{it}; \theta_{it})$. The value function depends on past emissions (the vector \bar{e}_{it}) for which compliance may not have occurred and on the current cost shock.³⁷ The facility's Bellman equation is then

³⁶ If the permits-to-emissions compliance ratio is 2:1, then the compliance factor is 0.5.

³⁷ The initial emissions vector is the zero vector, i.e., $\bar{e}_{i0} = 0$.

$$V_{it}(\bar{e}_{it}; \theta_{it}) = \min_{e_{it}d_{ijt}} c_{it}(e_{it}; \theta_{it}) + \sum_{t' \in \phi_i^{-1}(t)} \sum_j p_{jt} d_{ijt'} + \beta E_t V_{it+1}(\bar{e}_{it+1}; \theta_{it+1}) \quad (1)$$

subject to:

$$\bar{e}_{it+1}^t = \bar{e}_{it}^t + e_{it} \text{ for } t, \text{ and } \bar{e}_{it+1}^{t''} = \bar{e}_{it}^{t''} \text{ for } t'' \neq t \quad (2)$$

and

$$\bar{e}_{it+1}^{t'} \leq \sum_j \alpha_{ijt'} d_{ijt'} \text{ for every } t' \in \phi_i^{-1}(t) \quad (3)$$

The Bellman equation in Eq. 1 states that the optimal value of a given vector of past emissions and a realized cost shock is the minimized current abatement costs plus compliance costs plus the discounted expected optimized future value. The first term in the Bellman objective is the abatement cost. The second term in the Bellman is the compliance cost and is zero if period t is not a compliance period, i.e., if $\phi_i^{-1}(t)$ is empty. In a compliance period t , this term sums over all periods t' , which are “trued-up” in period t . For each of these periods, t' , the summation is over demands for all vintages. Note that the relevant permit price for compliance is the price at compliance time t . Even if the facility had purchased permits to cover its emissions obligation when the emissions occurred at time t' , as might be prudent, the relevant opportunity cost is the price of the permits at the time when they must be surrendered to cover the compliance obligation. The third term in the Bellman is the discounted, period t expectation of the optimized future value where E_t is the period t expectation operator.

Optimization of the Bellman in Eq. 1 is subject to two constraints. The constraint in Eq. 2 updates the t^{th} element of the past emissions vector so that the vector \bar{e}_{it+1} has period t' emissions as the t^{th} element but leaves every other element unchanged, i.e., $\bar{e}_{it+1}^{t'} = e_{it'}$, for $t' \leq t$ and has $\bar{e}_{it+1}^{t'} = 0$ for $t' > t$. The constraint in Eq. 3 ensures that demand for permits—weighted by compliance factors—is sufficient to cover the past emissions in the compliance period t .

The optimization in the Bellman can be simplified by noting that the objective and constraint are linear in the demands d_{ijt} . It is easy to show that optimality with

respect to d_{ijt} of Eq. 1 subject to Eq. 3 requires $\sum_j p_{jt} d_{ijt'} = \min_j \left\{ \frac{p_{jt}}{\alpha_{ijt'}} \right\} \bar{e}_{it+1}^{t'}$.³⁸

Intuitively, least cost compliance for period t' emissions requires choosing the cheapest applicable permits. Substituting this optimality equation into Eq. 1, shows that the optimization in Eq. 1-3 can be simplified to this Bellman:

$$V_{it}(\bar{e}_{it}; \theta_{it}) = \min_{e_{it}} c_{it}(e_{it}; \theta_{it}) + \sum_{t' \in \phi_i^{-1}(t)} \left\{ \min_j \left\{ \frac{p_{jt}}{\alpha_{ijt'}} \right\} \bar{e}_{it+1}^{t'} \right\} + \beta E_t V_{it+1}(\bar{e}_{it+1}; \theta_{it+1}) \quad (4)$$

subject to the constraint in Eq. 2.

Now consider optimization of this Bellman. If t is a compliance period, i.e., if $\phi_i(t) = t$, then $\bar{e}_{it+1}^t = e_{it}$, and the first order condition is:

$$-c'_{it}(e_{it}; \theta_{it}) = \min_j \left\{ \frac{p_{jt}}{\alpha_{ijt}} \right\} \quad (5)$$

Thus if period t is a compliance period, the firm sets the marginal abatement cost equal to the period t price of the cheapest valid permit.

Now consider the first order condition if t is not a compliance period, i.e., if $\phi_i(t) > t$. An increase in emissions does not lead to an immediate compliance liability, but rather increases the stock of past emissions for which compliance has not yet occurred (the relevant element of the vector \bar{e}_{it+1}) which affects $V_{it+1}(\bar{e}_{it+1}; \theta_{it+1})$. The first order condition can then be written:

$$\begin{aligned} -c'_{it}(e_{it}; \theta_{it}) &= \beta E_t \frac{\partial V_{it+1}}{\partial \bar{e}_{it+1}^t} \frac{\partial \bar{e}_{it+1}^t}{\partial e_{it}} = \dots = \beta^{\phi_i(t)-t} E_t \frac{\partial V_{i\phi_i(t)}}{\partial \bar{e}_{i\phi_i(t)}^t} \\ &= \beta^{\phi_i(t)-t} E_t \min_j \left\{ \frac{p_{j\phi_i(t)}}{\alpha_{ijt}} \right\} \end{aligned} \quad (6)$$

The first equality follows from optimization in period t , in which the marginal abatement cost is set equal to the discounted effect on the next period's value function through the effect on the past emissions vector, \bar{e}_{it+1} . Note that $\frac{\partial \bar{e}_{it+1}^t}{\partial e_{it}} = 1$ from Eq. 2. The second and third equalities follow since the increase in the past emissions vector has no effect on the subsequent value functions until the compliance period, $\phi_i(t)$. The final equality

³⁸ Cost minimization with a linear production function is analogous. In particular, if the linear production function is $f(k, l) = Ak + Bl$, then it is easy to show that the cost function is $c(q) = \min \left\{ \frac{r}{A}, \frac{w}{B} \right\} q$ where all variables follow the usual definitions.

follows from differentiation of Eq. 4 in the compliance period. Intuitively, Eq. 6 states that the facility sets marginal abatement costs equal to the period t expectation of the discounted price of the cheapest valid permit at the compliance time.

Eq. 5 and Eq. 6 characterize the facility's abatement/emissions decision—conditional on the price—for periods which are and are not compliance periods. To derive the demand for permits, we must go back to the d_{ijt} , which are the demands for the permits of each vintage.³⁹ Summing over time periods, $\sum_t d_{ijt}$, gives facility i 's demand for permits of vintage j . Summing over facilities, $\sum_i \sum_t d_{ijt}$, gives the market demand for permits of vintage j . Since the supply of each vintage of permits is fixed, the equilibrium must have $\sum_i \sum_t d_{ijt} = \bar{E}_j$ for each vintage j .

The equilibrium is not completely characterized by equating supply and demand, since there are time dated prices. Complete characterization of the equilibrium requires a condition on the relationship of prices across time, i.e., an arbitrage condition. Assume market participants (possibly speculators) would buy and hold all permits if $p_{jt} < \beta^{t'-t} E_t p_{jt'}$ and would not hold any permits if $p_{jt} > \beta^{t'-t} E_t p_{jt'}$. Thus vintage j permits will be both held and used by market participants in both periods only if $p_{jt} = \beta^{t'-t} E_t p_{jt'}$.

We can now state the first result:

Result 1: Compliance Invariance. If in the prompt compliance equilibrium for every t and t' with $t \leq t'$ we have that $p_{jt} = \beta^{t'-t} E_t p_{jt'}$ and $E_t \min_j \left\{ \frac{p_{jt'}}{\alpha_{ijt}} \right\} = \min_j \left\{ \frac{E_t p_{jt'}}{\alpha_{ijt}} \right\}$, then the equilibrium is invariant to delayed compliance.

Proof: To prove invariance, we show that the equilibrium price path with prompt compliance is also an equilibrium price path with delayed compliance. The sufficient conditions imply that:

$$\beta^{\phi_i(t)-t} E_t \min_j \left\{ \frac{p_{j\phi(t)}}{\alpha_{ijt}} \right\} = \min_j \left\{ \frac{\beta^{\phi_i(t)-t} E_t p_{j\phi(t)}}{\alpha_{ijt}} \right\} = \min_j \left\{ \frac{p_{jt}}{\alpha_{ijt}} \right\} \quad (7)$$

for the prompt compliance price path for any $\phi_i(t)$. For delayed compliance with this price path, marginal abatement costs would be set equal to the left-hand side of Eq. 7. Since marginal abatement costs would be set equal to the right-hand side of Eq. 7 with

³⁹ This derivation of d_{ijt} is presented in the proof of Result 1 and in Appendix A.

prompt compliance, Eq. 7 shows that e_{it} would be invariant to compliance timing if the price path is invariant. We show in Appendix A that the equilibrium price path is invariant since supply and demand are invariant to compliance timing. ■

The intuition of the result follows from comparing prompt compliance with delayed compliance. With prompt compliance, the facility sets its marginal abatement cost equal to the cheapest current price. With delayed compliance, the facility sets its marginal abatement cost equal to the cheapest expected future price. However, by the sufficient condition, the current price equals the expected future price. Thus the equilibrium is invariant to the delayed compliance.

The intuition of the result also follows from the cost-effectiveness of emissions trading. The emissions target of a cap and trade program is defined by the cap, i.e., by the permit supply and validity dates of permits. It is well known that cap and trade is cost effective, i.e., attains the emissions target at least cost.⁴⁰ Thus the prompt compliance equilibrium is cost effective. However, the delayed compliance equilibrium must also be cost effective and attains the same emissions target, i.e., delayed compliance does not affect the permit supply or validity dates of permits. Thus, the prompt and delayed compliance equilibria must be the same.

The first sufficient condition, $p_{jt} = \beta^{t'-t} E_t p_{jt'}$, states that the present value of the current price equals the present value of the expected future price. This condition, which is similar to the efficient markets hypothesis, would be enforced by intertemporal arbitrage by permit holders. Note that this sufficient condition might not hold, for example, if the regulator imposed price regulation or if the regulator injected or removed permits into the market. We return to these examples in Section 3.2.

The second sufficient condition, $E_t \min_j \left\{ \frac{p_{jt'}}{\alpha_{ijt}} \right\} = \min_j \left\{ \frac{E_t p_{jt'}}{\alpha_{ijt}} \right\}$, holds for most conceivable market designs. The condition need not hold in principle since the $\min\{\cdot\}$ function is concave and by Jensen's inequality $E_t \min_j \left\{ \frac{p_{jt'}}{\alpha_{ijt}} \right\} \leq \min_j \left\{ \frac{E_t p_{jt'}}{\alpha_{ijt}} \right\}$. However, for this inequality to be strict, it would require that one permit vintage would be cheapest for some realization of the cost shock, but that another permit vintage would be cheapest

⁴⁰ Cost effectiveness of emissions trading was first established by Montgomery (1972). Kling and Rubin (1997) show cost effectiveness, but not efficiency, in a dynamic setting.

for a different realization of the cost shock. This cannot happen if one vintage of permits is always more valuable than another substitute vintage, e.g., if permit validity is “nested”. With banking, for example, the earlier permits have wider applicability and can substitute for later permits, so the earlier permits are always more valuable. Even in RECLAIM where permit validity is not nested, when the abatement decision is made, the validity dates of the relevant permits are nested and the condition holds. In fact, we have been unable to construct a counter example where this condition fails to hold.

Although Result 1 shows that delayed compliance does not affect abatement costs, it does affect compliance costs. We present the second result as a corollary:

Corollary 1: Under the sufficient conditions of Result 1, the expected present value compliance costs are invariant to delayed compliance but the variance of compliance costs is higher with delayed compliance.

Proof: See Appendix A. ■

The intuition of Corollary 1 is quite straightforward. The sufficient condition $p_{jt} = \beta^{t'-t} E_t p_{jt'}$ implies that expectation of the present value of the current and future prices are equal, thus expected compliance costs are equal. Delaying compliance, however, means that compliance is settled at a later price, which is essentially the earlier price plus white noise. Thus the variance of compliance costs is higher, while the expected compliance costs are unchanged.

The invariance result and corollary are illustrated in Fig. 1 for a representative firm. The simplest illustration of delayed compliance timing involves two periods and a single vintage (vintage A) of permits, i.e., full banking and borrowing. Since $e_1 + e_2 = \bar{E}_A$, i.e., total emissions equal the cap, the example can be illustrated by measuring period 1 emissions from the left axis and period 2 emissions from the right axis where the width of the graph is \bar{E}_A . Marginal abatement costs for period 1 emissions, $-c'_1$, are downward sloping from the left axis and discounted marginal abatement costs for period 2 emissions, $-\beta c'_2$, are downward sloping from the *right* axis, i.e., upward sloping. The figure illustrates second period marginal abatement costs in present value and a discrete cost shock in the second period which can be either high ($-\beta c'_2{}^H$) or low ($-\beta c'_2{}^L$) with expected marginal abatement cost $-\beta E c'_2$.

In the prompt compliance equilibrium, since permits must be used in each period and held across both periods, the sufficient condition holds: $p_{A1} = \beta E p_{A2}$. By the first order conditions, we have $-c'_1(e_1) = p_{A1} = \beta E p_{A2} = \beta E c'_2(e_2)$, thus marginal abatement costs in the first period equal expected marginal abatement costs in the second period. Once period 1 emissions are chosen, emissions in period 2 simply exhaust the remaining permits and the price is determined by the marginal abatement cost as illustrated.

Since there is only a single vintage of permits used in both periods, both sufficient conditions hold, and the equilibrium is invariant to delayed compliance. With delayed compliance, the facility sets $-c'_1(e_1) = \beta E p_{A2}$. Since this gives exactly the same e_1 as with prompt compliance, the equilibrium is invariant.

Compliance costs, however, do depend on compliance timing. With prompt compliance, costs are either $p_{A1}e_1 + \beta p_{A2}^H e_2$ or $p_{A1}e_1 + \beta p_{A2}^L e_2$ depending on whether the cost shock is high or low. With delayed compliance, compliance costs are either $\beta p_{A2}^H \bar{E}_A$ or $\beta p_{A2}^L \bar{E}_A$ depending on whether the cost shock is high or low. Clearly the expected present value abatement costs are equal, but the variance is higher with delayed compliance since compliance for period 1 emissions has the additional white noise term $\beta p_{A2} - p_{A1}$.

Analysis of abatement cost shocks yields an additional insight about compliance timing: namely, equilibrium prices may not be unique and may be “degenerate”. Define *degenerate prices* as equilibrium prices which are not determined by sloping supply or demand curves. In this model, they are thus prices which are not determined by marginal abatement costs. Degenerate prices are problematic since a Walrasian tâtonnement process is unlikely to discover them.

To demonstrate non-unique and degenerate prices, Fig. 2 extends the example illustrated in Fig. 1 to include two permit vintages.⁴¹ Assume vintage A permits can be used in either period, but vintage B permits cannot be used in the first period, i.e., $\alpha_{A1} = \alpha_{A2} = \alpha_{B2} = 1$ and $\alpha_{B1} = 0$, e.g., banking with no borrowing. Note that vintage A permits have broader applicability and hence are more valuable for any cost shock. Assume all permits are allocated initially so that permits are held and used in both

⁴¹ Fig. 2 assumes that $\beta = 1$ so there is no discounting.

periods. The model thus satisfies both sufficient conditions and the equilibrium is invariant to compliance timing.

Fig. 2 illustrates the equilibrium for three different cost shock realizations in the first period. Panel A shows the equilibrium if the marginal abatement costs are high in period 1. In this equilibrium, the firm would like to use vintage B permits. Thus, it uses all the vintage A permits in the first period. In this case, markets are completely segregated so $e_1 = \bar{E}_A$ and $e_2 = \bar{E}_B$, and invariant prices are as illustrated and are defined by $p_{A1} = p_{A2}^H = p_{A2}^L$, and $p_{B1} = Ep_{B2}$.⁴²

For Panel A, the equilibrium vintage A prices are not unique in the second period. With prompt compliance, all vintage A permits are used in period 1, and no permits are held. Thus, any price pair p_{A2}^L and p_{A2}^H could be an equilibrium provided $p_{A2}^L \geq p_{B2}^L$ and $p_{A2}^H \geq p_{B2}^H$ and $p_{A1} \geq Ep_{A2}$, i.e., vintage B prices are always cheaper in the second period and no permits are held.⁴³ This non-uniqueness is irrelevant with prompt compliance, since supply and demand are zero in period 2. With delayed compliance, similar ambiguity is possible except now all permits must be held. In short, any price pair p_{A2}^L and p_{A2}^H could be an equilibrium provided $p_{A2}^L \geq p_{B2}^L$ and $p_{A2}^H \geq p_{B2}^H$ and $p_{A1} = Ep_{A2}$.⁴⁴ Unfortunately, non-uniqueness is relevant, since supply and demand are non-zero (equal to \bar{E}_A) in period 2 with delayed compliance. Note that any period 2, vintage A price not equal to p_{A1} is degenerate.

Panel B shows the equilibrium if the marginal abatement costs are somewhat lower in period 1. In this equilibrium, the firm would like to use vintage B permits in expectation. Thus, it uses all the vintage A permits in the first period, so $e_1 = \bar{E}_A$ and $e_2 = \bar{E}_B$. However, the period 2 markets cannot be completely segregated as in Panel A, i.e., $p_{A1} = p_{A2}^H = p_{A2}^L$ cannot be an equilibrium. This is because if abatement costs are high in the second period, period 2 marginal abatement costs are above those in period 1. Since the price of vintage B permits must rise in this contingency, the price of vintage A permits must also rise, otherwise demand for vintage A permits would exceed supply. Thus $p_{A2}^H \geq p_{B2}^H > p_{A1}$. Thus in any equilibrium with $p_{A1} = Ep_{A2}$, we must have prices

⁴² Invariant prices are prices which are equilibrium prices for both prompt and delayed compliance.

⁴³ Short selling would ensure that $p_{A1} = Ep_{A2}$, but would not guarantee uniqueness.

⁴⁴ We cannot have $p_{A1} < Ep_{A2}$, otherwise speculator demand for vintage A permits in period 1 would exceed supply.

p_{A2}^L and p_{A2}^H such that $p_{A2}^H \geq p_{B2}^H > p_{A1} = Ep_{A2} > p_{A2}^L \geq p_{B2}^L$. Since $Ep_{A2} > Ep_{B2}$, it is easy to see that $p_{A2}^H > p_{B2}^H$ or $p_{A2}^L > p_{B2}^L$. In other words, the equilibrium must have at least one degenerate price.

As in Panel A, the non-unique and degenerate period 2 prices are irrelevant with prompt compliance since supply and demand for vintage A permits are zero in period 2. However, with delayed compliance, the non-unique and degenerate period 2 prices are relevant since vintage A permits must be held until the second period.

Panel C shows the equilibrium if the marginal abatement costs are low in period 1. In this equilibrium, the firm would not want to borrow vintage B permits. In this case, markets are integrated; there is banking, $e_1 < \bar{E}_A$; and $e_2 > \bar{E}_B$; and the prices of vintage A and vintage B permits are equal in all contingencies. In fact, this is the equilibrium that would result with unlimited banking and borrowing as in Fig. 1.

We now state the second result:

Result 2: Non-unique and Degenerate Prices. The equilibrium prices may not be unique, and the equilibrium may require degenerate prices. These prices are relevant to the compliance decision iff compliance is delayed.

Proof: The possibility result is proved by the preceding example. ■

The example showing non-unique and degenerate prices is not an exceptional case in several respects. It does not require an extreme cost draw. It also does not require banking and similar examples could be constructed for borrowing. Moreover, the example does not require a discrete cost shock, and similar examples could be constructed for a continuously distributed cost shock.⁴⁵

The intuition of non-unique and degenerate prices can be explained by supply and demand in the permit market. In this model, supply is determined by the cap and hence is perfectly inelastic. Demand for permits is determined by marginal abatement costs and is elastic. However, once emissions occur, the demand for permits becomes perfectly inelastic.⁴⁶ With delayed compliance, both supply and demand are perfectly inelastic at the time of compliance, and hence any price can clear the market. Which price

⁴⁵ For this example with a continuous cost shock in Panel B, the price of vintage A permits in the second period would be distributed over an interval with positive density and a mass point on the lowest price in the interval. This mass point would be above the marginal abatement cost, i.e., would be degenerate.

⁴⁶ We assume throughout, quite reasonably (!), that emissions cannot be “unemitted” after they are emitted.

distribution is the actual equilibrium price distribution can depend on later abatement costs and the arbitrage condition. For degenerate prices, it is the arbitrage condition (rather than later abatement costs) which determines the degenerate prices.

3.2 Equilibria that vary with compliance timing

The invariance result is quite general and applies to a broad class of cap and trade markets. However, some market designs may not satisfy the sufficient conditions and hence the equilibrium may not be invariant. This subsection first shows that the equilibrium may vary if permit allocation is delayed such that permits are allocated after their first date of validity. Although not common in market designs, the EU-ETS and RGGI delay permit allocation. We then present an example with a price cap and reserve fund from Hasegawa and Salant (2010a, 2010b). This equilibrium is also not invariant to compliance timing. Several proposed cap and trade markets use price caps supported by reserve funds.

If permit allocation is delayed, the equilibrium may not be invariant since the supply of permits is not time invariant, and thus the current price may not equal the expected future price. This possibility is illustrated in Fig. 3. The simple two-period example in Fig. 3 assumes a single vintage of permits, which can be used in both periods, no cost shocks, and no discounting. The allocation of permits is delayed: \bar{E}_{A1} of the permits are allocated in the first period, and \bar{E}_{A2} of the permits are allocated in the second period. The cap is the total vintage A permits allocated $\bar{E}_A = \bar{E}_{A1} + \bar{E}_{A2}$. As illustrated in Fig. 3, the allocation of a relatively large proportion of the cap is delayed.

First consider the equilibrium with prompt compliance. The supply of permits at the first compliance time is \bar{E}_{A1} . Thus emissions in the first period cannot exceed \bar{E}_{A1} . As illustrated the first period allocation is relatively small, so first period emissions are \bar{E}_{A1} and second period emissions are \bar{E}_{A2} . The equilibrium prices are $p_{A1}^P = -c_1'(\bar{E}_{A1})$ and $p_{A2}^P = -c_2'(\bar{E}_{A2})$. Note that $-c_1'(\bar{E}_{A1}) > -c_2'(\bar{E}_{A2})$, so the facility would like to increase period 1 emissions and reduce period 2 emissions. However, this is not feasible

since the necessary permits are not available at compliance time. Note also that since $p_{A1}^P > p_{A2}^P$, the sufficient condition of the theorem does not hold.⁴⁷

Now consider the equilibrium with delayed compliance. In the compliance period the supply of permits is $\bar{E}_{A1} + \bar{E}_{A2}$. Since all permits are perfect substitutes, they trade at a common price p_{A2}^D . First period emissions are determined by the expectation of this price, and second period emissions are determined by this price so $-c'_1(e_1) = p_{A2}^D = -c'_2(e_2)$ and marginal abatement costs are equal. Since permits allocated in the first period must be held to the second period, $p_{A1}^D = p_{A2}^D$.

The example shows that delayed allocations can lead to different equilibria with prompt and delayed compliance. In addition, the delayed compliance equilibrium has lower abatement costs.⁴⁸

The Hasegawa-Salant example, which analyzes price collars, is motivated by Waxman-Markey. Waxman-Markey has a strategic reserve fund, which would hold quarterly auctions with a reserve price of \$28 adjusted for inflation. This reserve fund would act as a price ceiling at \$28. Other programs have similar price containment mechanisms. For example, RECLAIM did not have an explicit price containment mechanism but regulators were forced to intervene in 2001 to prevent the price from going too high. AB 32 has an “allowance price containment reserve” which releases additional permits in three tiers at prices of \$40, \$45, and \$50. RGGI allows increasing use of offsets when prices exceed trigger prices. Such price containment mechanisms are increasingly common in permit market design.

The invariance result may also fail in a market with a price cap supported by a reserve fund. To illustrate the Hasegawa-Salant example, consider a market that lasts T years, with a single vintage of permits, no cost shock, and an initial cap of E . The price ceiling is set at \bar{p} and is implemented by a reserve fund of X permits available at a price of \bar{p} .

⁴⁷ Short selling would not be possible, since there is no one from whom to borrow permits.

⁴⁸ It is worth noting, however, that the lower costs from delayed compliance only arise because the regulator failed to allocate sufficient permits in the first period and the shortage of available permits prevented cost-reducing trades. In fact, if sufficient permits were allocated in the first period, then the equilibrium would be invariant.

Compare delayed compliance in year T versus prompt (continuous) compliance. With delayed compliance all compliance obligations are settled at time T at the price p_T . With prompt compliance, obligations are settled at each time t at the prevailing price p_t . The price must grow at the rate of interest if permits are being held.

The equilibrium may or may not utilize the reserve fund of permits. For example, if permits are relatively abundant, permit demand from a price path growing at the rate of interest to \bar{p} may not exceed E . In this case, the reserve fund would be unused whether compliance is prompt or delayed, and the equilibrium would be invariant to compliance timing. However, if the initial cap were tighter, then the reserve fund would be used.

The solid line in Fig. 4 illustrates delayed compliance in a case where the reserve fund is used completely and the price grows from p_0^d to p_T^d which is above \bar{p} . Hasegawa and Salant point out the interesting dynamics of the reserve fund which is completely depleted in a speculative attack in year t_2 . Speculators are willing to buy permits at time t_2 and hold them until time T since they can sell them to regulated facilities at time T at price p_T^d . Regulated facilities recognize that they will be able to buy permits at time T at the price p_T^d , and that there will $E + X$ permits available. Thus abatement decisions are made based on the expected price at time T , so the price path p_0^d to p_T^d is the equilibrium price path.

With delayed compliance, it is irrelevant whether the facilities use the reserve fund permits to cover emissions from before or after t_2 since all emissions are trued up at time T . All that matters for the equilibrium price path is the total number of permits available at time T : $E + X$, i.e., the distribution of the permits between the initial cap and the reserve fund is immaterial.

With prompt compliance, Hasegawa and Salant point out that the distribution of the permits between the initial cap and the reserve fund can matter. Importantly, if the initial cap is small enough, then there may not be enough permits in the initial cap to cover demand for permits along the price path p_0^d to $p_{t_2}^d$. Suppose, for example, that the initial cap is exhausted at time t_1 along this price path. If compliance is delayed, the facility simply sets its marginal abatement cost equal to $p_{t_1}^d$. Although there are no permits available from the initial cap at time t_1 , the facility can safely wait until T to buy

the permits from the speculators.⁴⁹ With prompt compliance, the facility has no such option. It must procure the permits at time t_1 and the only place to get permits is from the reserve fund at price \bar{p} . But since $\bar{p} > p_{t_1}^d$, the marginal abatement cost would be less than the permit price, so the facility would optimally reduce emissions. Hasegawa and Salant show that in equilibrium the facilities bid up the price of the initial-cap permits.

The equilibrium for prompt compliance and a tight initial cap is illustrated by the dashed line in Fig. 4. The price path grows at the rate of interest from p_0^p to \bar{p} , at which time the initial-cap permits are exhausted. Facilities then purchase permits from the reserve fund at the price \bar{p} from time t_1 to t_3 and use them for immediate compliance. No speculator would be willing to hold permits from t_1 to t_3 since there is no capital gain on the permits. At t_3 there is a speculative attack on the reserve fund and speculators hold the permits or sell them for immediate use between t_3 and T . (See Hasegawa and Salant.)

Hasegawa and Salant also compare the efficiency of the two equilibria. Under the assumption that non-constant marginal damages are stationary, both equilibria are inefficient since they inefficiently delay abatement.⁵⁰ Note that the delayed compliance equilibrium delays abatement more than the prompt compliance equilibrium, i.e., delayed compliance exacerbates the welfare losses. In effect, prompt compliance restricts inefficient borrowing from the reserve fund. If on the other hand damages are constant or only depend on total emissions, then delaying abatement reduces costs. In this case, delayed compliance increases welfare.

4. Illustrating the model for common market designs

To illustrate the generality of the model, this section illustrates common market designs in the framework of the model. We first illustrate temporal restrictions on permit validity and then spatial restrictions on permit validity.

⁴⁹ Or buy the permits themselves at time t_2 .

⁵⁰ This inefficiency was first pointed out by Kling and Rubin (1997).

4.1 Temporal permit restrictions

We illustrate the model for four common restrictions on permit validity across time: banking, borrowing, overlapping vintages, and costly borrowing. Fig. 5 illustrates the equilibrium permit price across twelve quarters for the model with these four designs. By assuming that abatement costs are stationary (and certain) across the twelve quarters, abatement and emissions are captured completely by the permit price. In short, a higher permit price implies more abatement and lower emissions. We also assume that the regulator makes equal permit allocations across vintages, i.e., the regulator is not trying to reduce emissions over time but rather to hold emissions at a desired level. Thus we are illustrating the steady-state of the market perhaps after some transition phase in which permits (and hence emissions) decrease over time.

Each panel of Fig. 5 illustrates a different market design in terms. In each panel, the permit price for each quarter is illustrated if the permit is valid for that quarter.⁵¹ The permit price is circled if there is positive demand for the permit at that price (i.e., if that vintage of permits is used for that quarter's compliance obligation). Only the cheapest permits are used since permits are perfect substitutes. All prices grow at the rate of interest as required by the arbitrage condition.

Panel A of Fig. 5 illustrates perhaps the most common market design: banking without borrowing as in ARP, CAIR, and CSAPR. To illustrate the design, we show three vintages of permits with equal allocations. Vintage A permits are allocated for year 1 so are valid for all four quarters of year 1, and—since they can be banked—are also valid in years 2 and 3, i.e., $\alpha_{iA1} = \alpha_{iA2} = \alpha_{iA3} = 1$. Since Vintage A permits are valid in all three years, the price is illustrated for all three years. Similarly, Vintage B permits can be used in years 2 and 3, i.e., $\alpha_{iB1} = 0$ and $\alpha_{iB2} = \alpha_{iB3} = 1$, and Vintage C permits can only be used in year 3, i.e., $\alpha_{iC1} = \alpha_{iC2} = 0$ and $\alpha_{iC3} = 1$. Since Vintage A permits can be used in every quarter where Vintage B permits can be used, their price must be higher. Similarly, Vintage B permits are more valuable than Vintage C permits. Thus $p_{At} \geq p_{Bt} \geq p_{Ct}$ for every t .

⁵¹ Although equilibrium prices are defined for all quarters—since prices grow at the rate of interest—we simply illustrate the prices that are relevant for the abatement decision in each quarter.

To develop the intuition of the equilibrium, first consider the equilibrium if the permits could not be banked or borrowed. Since permit price grows at the rate of interest within each year, abatement is delayed within the year. However, abatement cannot be delayed across years. In fact, since permit allocations are equal and abatement costs are stationary, the marginal abatement costs and equilibrium permit prices must be equal across years, i.e., $p_{At} = p_{Bt+4} = p_{Ct+8}$.

Now consider what happens when the permits can be banked. The firm now could use Vintage A permits in year 2, but it would not, since their price is higher. Thus, the equilibrium with banking, illustrated in Panel A, is identical to the equilibrium without banking: abatement is delayed within each year; each vintage of permits is used exclusively in its allocation year; and no permits are banked.⁵²

Panel B of Fig. 5 shows an alternative market design where permits can be freely borrowed but not banked. This is similar to the three-year borrowing allowed in RGGI and AB 32.⁵³ Now all three vintages can be used in the first year since Vintages B & C can be borrowed. Thus Vintage C permits are the most valuable and $p_{At} \leq p_{Bt} \leq p_{Ct}$ for all t . Allowing borrowing *does* change the equilibrium since firm would like to borrow permits and delay abatement further. In equilibrium, borrowing drives down the price of Vintage A permits, but increases the price of Vintage C as illustrated in Panel B. Thus abatement is delayed across years; permits are borrowed; and the permit prices for all vintages are equal.⁵⁴

Panel C shows a market design adopted by the RECLAIM program in southern California. The model has overlapping permit vintages each of which is valid for four quarters, with new permits being issued and old permits expiring every two quarters. This equilibrium is quite similar to the equilibrium with banking since each vintage of permits is exhausted before their expiration date when new permits become valid. Here, the more frequent allocation of permits leads to less delayed abatement within a year.

Panel D illustrates a feature introduced by the Waxman-Markey proposal. Under the proposal, the permits can be freely banked as in Panel A. However, the permits can

⁵² Since there are no abatement cost shocks and the supply of permits is not decreasing, there is no reason to bank permits for the future.

⁵³ RGGI and AB 32 also allow banking. However, banking does not occur in this simple equilibrium.

⁵⁴ This equilibrium with borrowing is equivalent to the equilibrium with banking and borrowing.

also be borrowed, but at a cost: an 8% quantity-based penalty is imposed on each permit for each year it is used before its allocation date.⁵⁵ Thus the “effective price” of a Vintage B permit is higher if it is used in year 1 rather than in year 2. These effective prices are illustrated in Panel D. As illustrated, the borrowing penalty is sufficient that no permits are borrowed in equilibrium. However, if the borrowing penalty were less severe (or if there were adverse abatement cost shocks) firms might borrow permits in equilibrium even though borrowing is costly.

4.2 Spatial permit restrictions

To illustrate spatial validity in permit design, consider two facilities with different (constant) marginal damages, $\delta_1 < \delta_2$ and abatement costs $c_1(e_1)$ and $c_2(e_2)$. Marginal abatement costs, damages, and efficient emissions are illustrated in Fig. 6. We illustrate how this efficient allocation can be implemented with two different spatial restrictions on permit validity.

We first illustrate zonal validity of two vintages of permits. This example is similar to the RECLAIM program where inland facilities can use either coastal or inland permits, but coastal facilities can only use coastal permits. Define two spatial vintages of permits: A and B. The low damage facility (Facility 1) can use either vintage of permits whereas the high damage facility (Facility 2) can only use the vintage B permits. Since vintage B permits can be used by either facility, they are more valuable and $p_A \leq p_B$. The spatial restrictions are then $\alpha_{1A} = \alpha_{1B} = 1$, $\alpha_{2A} = 0$, and $\alpha_{2B} = 1$.

To implement the efficient allocation in Fig. 6 with this market design, the regulator simply chooses the caps such that $\bar{E}_A = e_1$ and $\bar{E}_B = e_2$. With these caps, the two vintages are segregated in equilibrium with $p_A = \delta_1 < \delta_2 = p_B$, and emissions are at the efficient levels. Facility 2 would like to use the vintage A permits but cannot due to the spatial restriction. Facility 1 is allowed to use the vintage B permits but chooses not to since they are more expensive.

Next consider implementing the equilibrium with one spatial vintage of permits (vintage C) but different permits-to-emissions compliance ratios for the two facilities.

⁵⁵ Using our notation, $\alpha_{iA1} = \alpha_{iA2} = \alpha_{iA3} = 1$; $\alpha_{iB1} = 1/1.08 = 0.93$ and $\alpha_{iB2} = \alpha_{iB3} = 1$; and $\alpha_{iC1} = 0.86$, $\alpha_{iC2} = 0.93$, and $\alpha_{iC3} = 1$.

The regulator now chooses the compliance ratios: α_{1C} and α_{2C} , and the cap: \bar{E}_C , to implement the efficient allocation in Fig. 6.⁵⁶

If the regulator sets compliance ratios $\alpha_{1C} = 1$ and $\alpha_{2C} = \delta_1/\delta_2 < 1$ and the cap $\bar{E}_C = e_1 + e_2/\alpha_{2C}$, then the equilibrium price is $p_C = \delta_1$. Facility 1 then sets its marginal abatement cost equal to $p_C/\alpha_{1C} = p_C = \delta_1$, and Facility 2 sets its marginal abatement cost equal to $p_C/\alpha_{2C} = \delta_2\delta_1/\delta_2 = \delta_2$. Since Facility 1 demands e_1 permits and facility 2 demands e_2/α_{2C} permits, the equilibrium implements the efficient emissions levels.

These two market designs each implement the efficient spatial distribution of emissions with different mechanisms and are equivalent in this simple model. However, they would not be equivalent with abatement cost shocks. For example, if Facility 1 suffered an adverse abatement cost shock, the zonal design would require that only Facility 1 adjusts, i.e., only p_A increases, unless the abatement cost were so extreme that p_A were driven above δ_2 , at which point $p_A = p_B$ and both facilities would reduce emissions. On the contrary, with different permits-to-emissions compliance ratios, even a small abatement cost shock to Facility 1 causes p_C to increase which causes both facilities to reduce emissions. Which market design is more efficient depends on the relative slopes of marginal abatement costs and marginal damages.⁵⁷

5. Conclusion

Market mechanisms, primarily cap and trade programs, continue to be one of the primary tools for reducing harmful emissions. To address a variety of technical and political constraints, existing and proposed programs take a variety of approaches to permit validity and compliance timing. We review these approaches for nine major cap and trade programs. ARP defined the classic program design, and several of its features have been widely adopted: annual permits with banking but no borrowing, 1:1 permits-to-emissions compliance ratios, quarterly reporting, and annual compliance. However, as shown in Section 2, other programs depart in significant ways from the design of the

⁵⁶ The differential permits-to-emissions compliance ratios in CAIR and ARP are similar to this market design. Under CAIR, SO₂ permit allocations in the eastern United States have a 2:1 compliance ratio for 2010 to 2014, increasing to 2.86:1 after 2014. SO₂ permits under ARP continue to cover emissions at a 1:1 rate outside the designated eastern states.

⁵⁷ A more complete analysis of optimal spatial design with asymmetric information is beyond the scope of this paper.

ARP. One significant departure, adopted by RGGI and AB32, is the three-year compliance window. Other departures impose different compliance ratios, restricted forms of borrowing, or restricted forms of banking. Despite this considerable heterogeneity, there has been little systematic study of these key features of program design.

Our model of permit market design provides a general analysis of permit validity and compliance timing. The first result shows that equilibrium abatement is invariant to compliance timing if a simple arbitrage condition holds. If abatement is invariant to compliance timing, delayed compliance cannot smooth adverse cost shocks. In particular, delayed compliance does nothing to help regulated facilities respond to adverse shocks, for example, from high electricity demand due to adverse weather, from fuel price spikes, or from clean generator outages.

Despite the invariance results, our second and third results suggest caution in adopting delayed compliance due to the increased variance of compliance costs and to non-unique or degenerate prices. With delayed compliance, compliance obligations are settled at the expected price plus noise. Thus the variance of compliance costs increases with delayed compliance. Furthermore, there may be no elasticity in the supply of or demand for permits at the time of compliance, and hence any price could clear the market. Which prices form the equilibrium price distribution is determined by both future abatement cost shocks and the arbitrage condition. Degenerate prices are those determined by the arbitrage condition rather than abatement costs, i.e., degenerate prices are prices that *ex post* justify the *ex ante* expectation. These degenerate prices, which would likely not be discovered by a Walrasian tâtonnement process, are irrelevant with prompt compliance but govern transactions with delayed compliance.

Although the invariance results are quite general, there are some program designs which do not satisfy the sufficient conditions of the results. In particular, delayed allocations or a price cap supported by a reserve fund can lead to abatement that depends on compliance timing.

The equilibrium of the model is illustrated using several specific features of temporal and spatial permit validity. These features include banking, borrowing, borrowing with interest, spatial segmentation, and permits-to-emissions compliance

ratios. The illustrations demonstrate the model's broad applicability to market designs in practice.

The increasing heterogeneity in program design requires careful study of different aspects of the programs. Our model and analysis present a framework for analyzing permit validity and compliance timing. The invariance results show that compliance timing cannot smooth abatement cost shocks for a broad class of models. In these cases, moreover, delayed compliance increases the variance of compliance costs and can imply non-unique or degenerate equilibrium prices. Finally, several other factors would also be assessed when considering prompt versus delayed compliance. These include effects on program administration (administrative costs, smoothing the flow of administrative work, developing/maintaining staff expertise, prompt resolution of disputes and data errors, and maintaining the credibility of the program) and market performance (improving permit supply information, increasing the salience of compliance costs, and avoiding complications from bankruptcy). These factors seem to favor prompt compliance, and they can be considered along with our results when designing cap and trade programs.

References

- Bushnell, James B., Howard Chong, and Erin T. Mansur. 2009. "Profiting from Regulation: An Event Study of the EU Carbon Market." NBER Working Paper No. 15572
- California Environmental Protection Agency, Air Resources Board. 2010. Volume 1: Proposed Regulation to Implement the California Cap-and-Trade Program; Appendix A: Proposed Regulation Order, posted October 28, 2010. <http://www.arb.ca.gov/regact/2010/capandtrade10/capv1appa.pdf>.
- Carlson, Dale, Charles Forman, Nancy Olmstead, John Ledyard, Charles Plott, David Porter and Anne Sholtz. 1993. "An Analysis and Recommendation for the Terms of the RECLAIM Trading Credit," Technical Report, South Coast Air Quality Management District, April 1993.
- Carlson, Dale A. and Anne M. Sholtz. 1994. Designing Pollution Market Instruments: Cases of Uncertainty. *Contemporary Economic Policy* 12: 114-125.
- Convery, Frank J. and Luke Redmond. 2007. "Market and Price Developments in the European Union Emissions Trading Scheme" *Review of Environmental Economics and Policy* 1 (1): 88-111.
- Ellerman, A. Denny, and Paul L. Joskow. 2008. "The European Union's Emissions Trading System in Perspective." Pew Center on Global Climate Change. May.
- Ellerman, A. Denny, Paul L. Joskow, and David Harrison, Jr. 2003. "Emissions Trading in the U.S.: Experience, Lessons, and Considerations for Greenhouse Gases." Pew Center on Global Climate Change. May.
- Ellerman, A. Denny, Paul L. Joskow, Richard Schmalensee, Juan-Pablo Montero, and Elizabeth M. Bailey. 2000. *Markets for Clean Air: The U.S. Acid Rain Program*. Cambridge, U.K.: Cambridge University Press.
- Ellerman, A. Denny and Raphael Trotignon. 2009. "Cross Border Trading and Borrowing in the EU ETS," *The Energy Journal*, International Association for Energy Economics, 30(Special I): 53-78.
- European Commission. 2003. Directive 2003/87/EC of the European Parliament and of the Council of 13 October 2003 Establishing a Scheme for Greenhouse Gas Emission Allowance Trading Within the Community and Amending Council Directive 96/61/EC," Official Journal of the European Union, L 275/32 , 25.10.2003.
- Federal Register. 2005. Environmental Protection Agency: Rule to Reduce Interstate Transport of Fine Particulate Matter and Ozone (Clean Air Interstate Rule);

- Revisions to Acid Rain Program; Revisions to the NO_x SIP Call; Final Rule. Vol. 70, No. 91. May 12.
- Fowlie, Meredith and Nicholas Muller. 2010. “Designing markets for pollution when damages vary across sources : Evidence from the NO_x Budget Program.” mimeo.
- Hasegawa, Makoto and Stephen Salant. 2010a. “The Proposed Cap-and-Trade Program to Limit Greenhouse Gas Emissions: The Case of the Unbuttoned Collar” mimeo.
- Hasegawa, Makoto and Stephen Salant. 2010b. “The Proposed Cap-and-Trade Program Under Annual Compliance” mimeo.
- Holland, Stephen P. and Michael R. Moore. 2012. “When to Pollute, When to Abate? Intertemporal Permit Use in the Los Angeles NO_x Market” *Land Economics* 88(2): 275-299.
- Joskow, Paul L., Richard Schmalensee, and Elizabeth M. Bailey. 1998. “The Market for Sulfur Dioxide Emissions.” *American Economic Review* 88(4): 669-685.
- Kling, Catherine and Jonathan Rubin. 1997. “Bankable Permits for the Control of Environmental Pollution.” *Journal of Public Economics* 64(1): 101-115.
- Kruger, Joseph A., and William A. Pizer. 2004. “Greenhouse Gas Trading in Europe: The New Grand Policy Experiment.” *Environment* 46(8): 8-23.
- Montgomery, David. 1972. “Markets in Licenses and Efficient Pollution Control Programs.” *Journal of Economic Theory*, 5: 395-418.
- Muller, Nicholas Z., and Robert Mendelsohn. 2009. “Efficient Pollution Regulation: Getting the Prices Right.” *American Economic Review*, 99(5): 1714-1739.
- Regional Greenhouse Gas Initiative. 2007. Overview of RGGI CO₂ Budget Trading Program, http://rggi.org/docs/program_summary_10_07.pdf, accessed September 13, 2009.
- Regional Greenhouse Gas Initiative. 2008. Regional Greenhouse Gas Initiative Model Rule, <http://www.rrgi.org/docs/Model%20Rule%20Revised%2012.31.08.pdf>, accessed September 13, 2009.
- Schackenbach, John, Robert Vollaro, and Reynaldo Forte. 2006. “Fundamentals of Successful Monitoring, Reporting, and Verification under a Cap-and-Trade Program” *Journal of Air & Waste Management Association*. 56:1576–1583
- Schennach, Susanne. 2000. “The Economics of Pollution Permit Banking in the Context of Title IV of the 1990 Clean Air Act Amendments.” *Journal of Environmental Economics and Management* 40: 189-210.

- South Coast Air Quality Management District. 2009. Regulation XX: Regional Clean Air Incentives Market (RECLAIM), http://www.aqmd.gov/rules/reg/reg20_tofc.html, accessed December 1, 2009.
- Stavins, Robert N. 1998. "What Can We Learn from the Grand Policy Experiment? Lessons from SO₂ Allowance Trading." *Journal of Economic Perspectives* 12(3): 69-88.
- USEPA. 2002. "An Evaluation of the South Coast Air Quality District's Regional Clean Air Incentives Market – Lessons in Environmental Markets and Innovation." November <http://www.epa.gov/region09/air/reclaim/report.pdf>.
- USEPA. 2005a. "NO_x Budget Trading Program Progressive Flow Control." Accessed November 17, 2009 at <http://www.epa.gov/airmarkets/progsregs/nox/docs/flowcontrol.pdf>.
- USEPA. 2005b. "Evaluating Ozone Control Programs in the Eastern United States: Focus on the NO_x Budget Trading Program, 2004" EPA454-K-05-001.
- USEPA. 2006. "An Overview of the Regional Clean Air Incentives Market (RECLAIM)," Staff Paper, EPA Clean Air Markets Division, August 14.
- USEPA. 2009a. Title IV of the 1990 Amendments to the Clean Air Act: Acid Deposition Control. Accessed on October 13, 2009 at <http://www.epa.gov/airmarkt/progsregs/arp/docs/title4.pdf>.
- USEPA. 2009b. Acid Rain Program SO₂ Allowances Fact Sheet. Accessed October 13, 2009 at <http://www.epa.gov/airmarkt/trading/factsheet.html>.
- USEPA. 2009c. NO_x Budget Trading Program – Basic Information. Accessed November 17, 2009 at <http://www.epa.gov/airmarket/progsregs/nox/docs/NBPbasicinfo.pdf>.
- USEPA. 2011. Federal Implementation Plans: Interstate Transport of Fine Particulate Matter and Ozone and Correction of SIP Approvals. *Federal Register*, Vol. 76, No. 152, August 8, 2011.
- U.S. House of Representatives. 2009. American Clean Energy and Security Act of 2009; Title VII: Global Warming Pollution Reduction Program, passed by the U.S. House of Representatives. Accessed on December 1, 2009 at < <http://thomas.loc.gov/cgi-bin/query/D?c111:3:./temp/~c111om8fov::>>.
- Yates, Andrew J. and Mark B. Cronshaw. 2001. "Pollution Permit Markets with Intertemporal Trading and Asymmetric Information." *Journal of Environmental Economics and Management* 42, 104–118.

Appendix A

Proof of claim in Result 1

To show that the equilibrium price path is invariant to compliance timing, consider the demands. With prompt compliance, demand d_{ijt} is the open interval $[0, e_{it}]$ if $\frac{p_{jt}}{\alpha_{ijt}} = \min_j \left\{ \frac{p_{jt}}{\alpha_{ijt}} \right\}$ and zero otherwise. With delayed compliance, demand d_{ijt} is the open interval $[0, e_{it}]$ if $\frac{p_{j\phi(t)}}{\alpha_{ijt}} = \min_j \left\{ \frac{p_{j\phi(t)}}{\alpha_{ijt}} \right\}$ and zero otherwise. We show below that $\frac{p_{jt}}{\alpha_{ijt}} = \min_j \left\{ \frac{p_{jt}}{\alpha_{ijt}} \right\}$ implies that $\frac{p_{j\phi(t)}}{\alpha_{ijt}} = \min_j \left\{ \frac{p_{j\phi(t)}}{\alpha_{ijt}} \right\}$ under the sufficient conditions. Thus demands d_{ijt} are invariant for every i, j , and t . Since permit supply does not depend on timing, supply and demand are invariant and hence the equilibrium price path is invariant as well.

To show that $\frac{p_{jt}}{\alpha_{ijt}} = \min_j \left\{ \frac{p_{jt}}{\alpha_{ijt}} \right\}$ implies that $\frac{p_{j\phi(t)}}{\alpha_{ijt}} = \min_j \left\{ \frac{p_{j\phi(t)}}{\alpha_{ijt}} \right\}$, assume not, i.e., assume $\frac{p_{j\phi(t)}}{\alpha_{ijt}} > \min_j \left\{ \frac{p_{j\phi(t)}}{\alpha_{ijt}} \right\}$. Now taking period t expectations implies $E_t \frac{p_{j\phi(t)}}{\alpha_{ijt}} > E_t \min_j \left\{ \frac{p_{j\phi(t)}}{\alpha_{ijt}} \right\} = \min_j \left\{ \frac{E_t p_{j\phi(t)}}{\alpha_{ijt}} \right\}$ where the inequality comes from taking expectations and the equality follows by the sufficient condition of the theorem. But since $E_t p_{j\phi(t)} = \beta^{t-\phi(t)} p_{jt}$, this implies that $\frac{p_{jt}}{\alpha_{ijt}} > \min_j \left\{ \frac{p_{jt}}{\alpha_{ijt}} \right\}$ which is a contradiction.

Proof of Corollary 1

The present value of compliance costs with prompt compliance for facility i is $\sum_t \sum_j \beta^t p_{jt} d_{ijt}$. With delayed compliance the compliance costs are:

$$\sum_t \sum_j \beta^{\phi(t)} p_{j\phi(t)} d_{ijt} = \sum_t \sum_j \beta^t p_{jt} d_{ijt} + \sum_t \sum_j (\beta^{\phi(t)} p_{j\phi(t)} - \beta^t p_{jt}) d_{ijt} \quad (8)$$

$\equiv X + Y$

where X and Y are defined in Eq. 8. Since the prices and demands are invariant to the compliance timing, the first term on the RHS of Eq. 8, X , is the compliance cost with prompt compliance. Thus to prove that expected compliance costs are invariant, we prove that $EY = 0$. To prove that the variance is higher, we show that the covariance of

the two terms is zero since $Cov(X, Y) = 0$ implies that $Var(X + Y) = Var(X) + Var(Y)$.

$EY = 0$ follows directly since:

$$E \sum_t \sum_j (\beta^{\phi(t)} p_{j\phi(t)} - \beta^t p_{jt}) d_{ijt} = E \sum_t \sum_j (E_t \beta^{\phi(t)} p_{j\phi(t)} - \beta^t p_{jt}) d_{ijt} = 0$$

where the first equality follows from the expectation operator and the second equality follows since $p_{jt} = \beta^{t'-t} E_t p_{jt'}$.

Since $EY = 0$, it follows that $Cov(X, Y) = EXY$. Since X and Y are both sums, their product contains many terms. Letting j' and t' index the term from X , and letting j and t index the term from Y , the expectation of each term of the product is

$$E \beta^{t'} p_{j't'} d_{ij't'} (\beta^{\phi(t)} p_{j\phi(t)} - \beta^t p_{jt}) d_{ijt} = E \beta^{t'} p_{j't'} d_{ij't'} (E_t \beta^{\phi(t)} p_{j\phi(t)} - \beta^t p_{jt}) d_{ijt}$$

But this equals zero since $p_{jt} = \beta^{t'-t} E_t p_{jt'}$. Thus $Cov(X, Y) = 0$.

The model without uncertainty

This appendix presents the main model presented in Section 3 without the abatement costs shocks. This simplifies the analysis significantly by allowing Lagrangian techniques instead of dynamic programming. The presentation here also focuses more carefully on deriving demands for the permits and characterizing the competitive equilibrium. All results from the more general model in Section 3 hold. All notation is as defined in Section 3 with the exception that $c_{it}(e_{it})$ is the abatement cost function which is not subject to shocks.

Each facility minimizes compliance costs by choosing emissions in each period e_{it} and how much of each permit to use in each period, d_{ijt} . Facility i 's cost minimization problem is now

$$\min_{e_{it} d_{ijt}} \sum_{t=1}^{\infty} \left\{ \beta^t c_{it}(e_{it}) + \beta^{\phi_i(t)} \sum_{j=1}^{\infty} p_{j\phi_i(t)} d_{ijt} \right\} \quad (9)$$

subject to

$$e_{it} \leq \sum_{j=1}^{\infty} \alpha_{ijt} d_{ijt} \quad (10)$$

The objective in [10] minimizes the sum of abatement costs and permit costs subject to the compliance constraint in [11]. There are several things to note about the optimization. First, the abatement costs are incurred at the time the emissions take place. We assume (quite reasonably!) that after the pollution is emitted, there is no way for the facility to reduce its past emissions. Thus the relevant discount factor for abatement costs is β^t . Second, the relevant price of the permits is the price at the time of compliance, $p_{j\phi_i(t)}$. Even if the facility purchases the permits earlier (as would be prudent), the relevant cost is the opportunity cost at the time of compliance. Thus the relevant discount factor for compliance costs is $\beta^{\phi_i(t)}$. Third, the time of compliance may differ across facilities. Finally, a facility is in compliance if emissions in period t are less than the permits (weighted by compliance factors) summed over all j .

Assuming the constraint binds with equality, the Kuhn-Tucker first order and complementary slackness (C.S.) conditions for the facility's cost minimization are then:

$$d_{ijt} \geq 0 \quad \alpha_{ijt} \beta^t c'_{it}(e_{it}) + \beta^{\phi_i(t)} p_{j\phi_i(t)} \geq 0 \quad \text{C.S.} \quad (11)$$

for every i, j , and t . This condition states that if demand for a vintage of permits is positive, i.e., if $d_{ijt} > 0$, then the marginal abatement cost and weighted permit price must be equal in present value, i.e., $-\beta^t c'_{it}(e_{it}) = \beta^{\phi_i(t)} \frac{p_{j\phi_i(t)}}{\alpha_{ijt}}$. More generally, the marginal abatement cost must be less than or equal to the permit price (in present value), i.e., $-\beta^t c'_{it}(e_{it}) \leq \beta^{\phi_i(t)} \frac{p_{j\phi_i(t)}}{\alpha_{ijt}}$ which implies

$$-\beta^t c'_{it}(e_{it}) = \beta^{\phi_i(t)} \min_j \left\{ \frac{p_{j\phi_i(t)}}{\alpha_{ijt}} \right\} \quad (12)$$

if emissions are positive. The condition in [13] states that marginal abatement costs in each period should be equal (in present value) to the minimum weighted price of all the vintages that are valid for period t .⁵⁸

The cost minimization problem takes prices as given. Equilibrium prices will equate supply and demand in every period. Let \bar{p} be the vector of all time dated prices for all permit vintages. The cost minimization defines a demand correspondence for each vintage of permit. Let $D_{ij}(\bar{p})$ be firm i 's demand correspondence for permit vintage j . Since permits of different vintages are perfect substitutes, $D_{ij}(\bar{p})$ will generally not be a function. Clearly, $D_{ij}(\bar{p}) = \sum_{t=1}^{\infty} d_{ijt}$, i.e., facility i 's demand for each permit vintage is the sum over the demands for all the periods the vintage is valid. Since market demand is found by adding up demands from each facility, and since permit supply is perfectly inelastic, market demand equals supply for permit vintage j if:

$$\sum_i D_{ij}(\bar{p}) = E_j. \quad (13)$$

Describing the market demand and supply for permits would normally characterize the competitive equilibrium. However, prices are time dated, so there are more prices than markets. Since permits are costless to store, arbitrage will force the permit prices to be equal in present value, i.e., to grow at the rate of interest. Let p_{j0} be this common present value, i.e., $p_{j0} \equiv \beta^t p_{jt}$ for every t , which characterizes the arbitrage condition. This condition reduces the dimensionality of the price vector to the dimension of the number of markets, and the equilibrium is now completely characterized.

Using the arbitrage condition, equation [13] can now be written

$$-\beta^t c'_{it}(e_{it}) = \min_j \left\{ \frac{p_{j0}}{\alpha_{ijt}} \right\} \quad (14)$$

or equivalently

⁵⁸ As in Section 3, [4] can be derived by noting that [1] is linear in the d_{ijt} controls. The optimal solution is then a corner solution at the minimum of the prices. After substitution of the binding constraint, [4] is the FOC for e_{it} .

$$-c'_{it}(e_{it}) = \min_j \left\{ \frac{p_{jt}}{\alpha_{ijt}} \right\} \quad (15)$$

Note that [16], which implies that the facility sets the marginal abatement cost in period t equal to the cheapest valid permit in period t , is equivalent to [7] in Section 3. Thus the current permit prices contain all the relevant information for the facility to make its abatement cost decisions.

Figures

Figure 1: Invariance with two periods and one vintage of permits

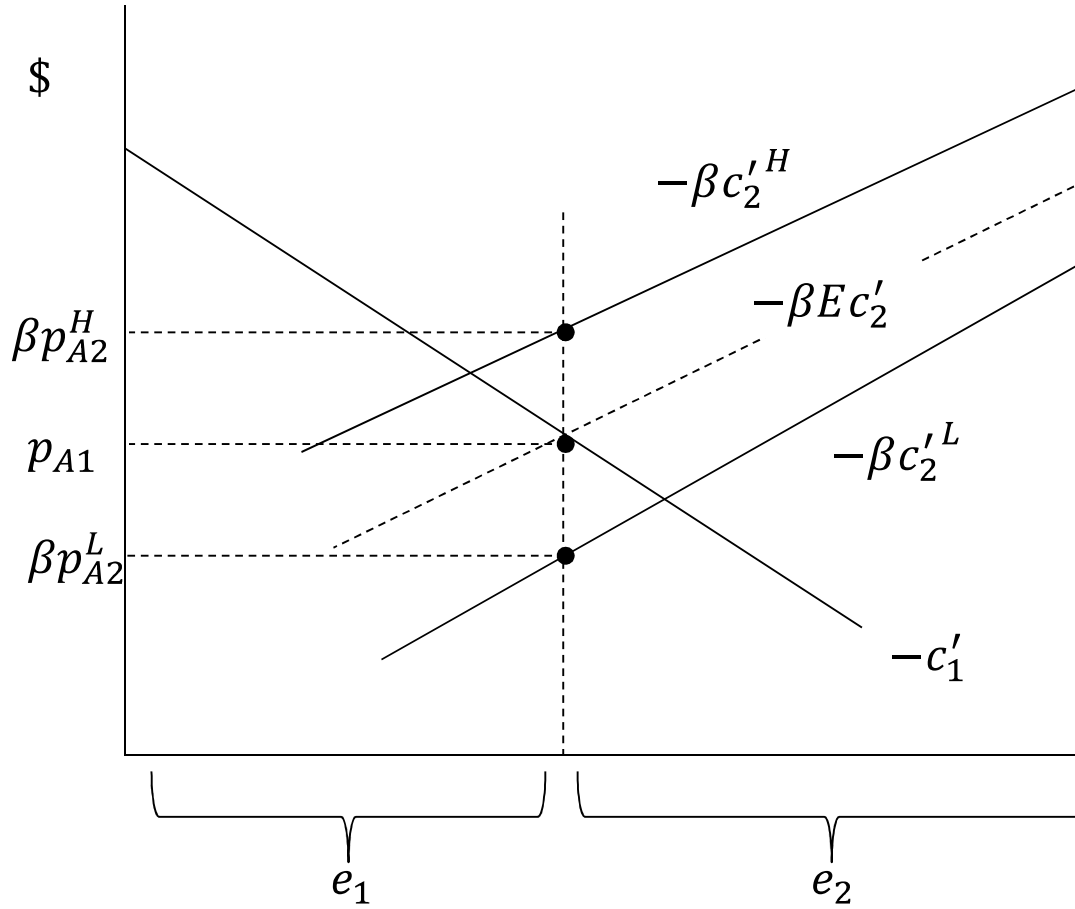
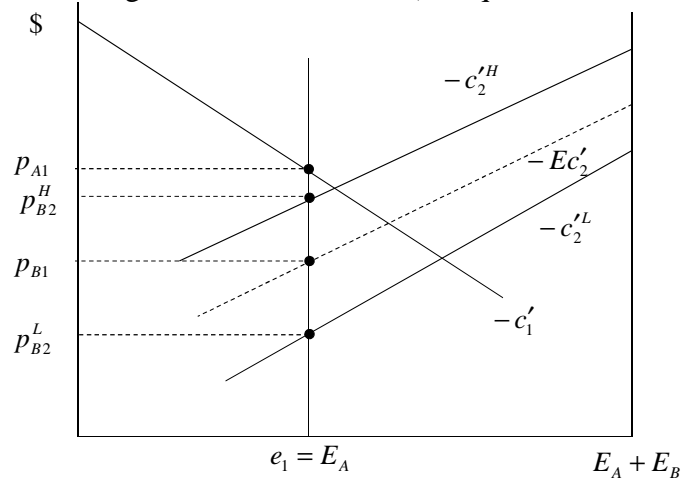
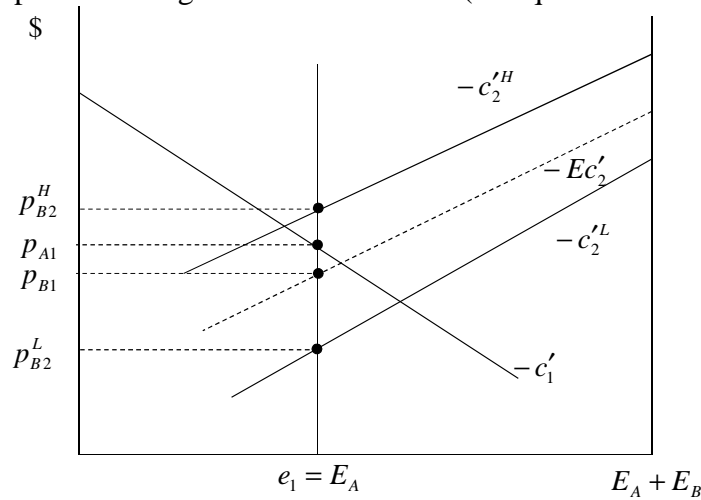


Figure 2: Invariance with Non-unique and Degenerate Prices

Panel A: High period 1 marginal abatement costs (no equilibrium banking)



Panel B: Medium period 1 marginal abatement costs (no equilibrium banking)



Panel C: Low period 1 marginal abatement costs (equilibrium banking)

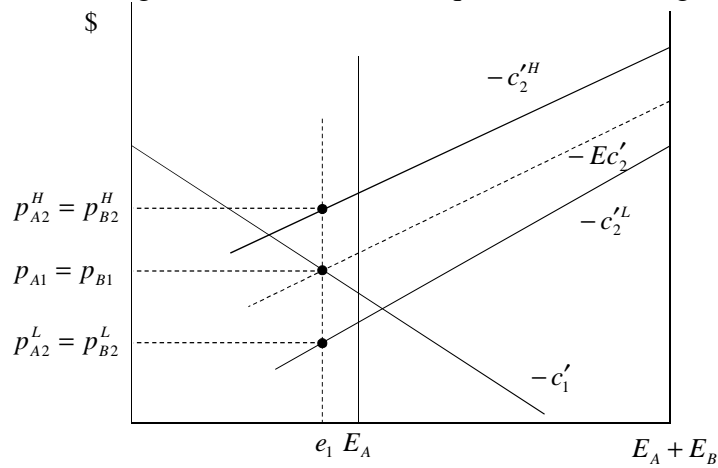
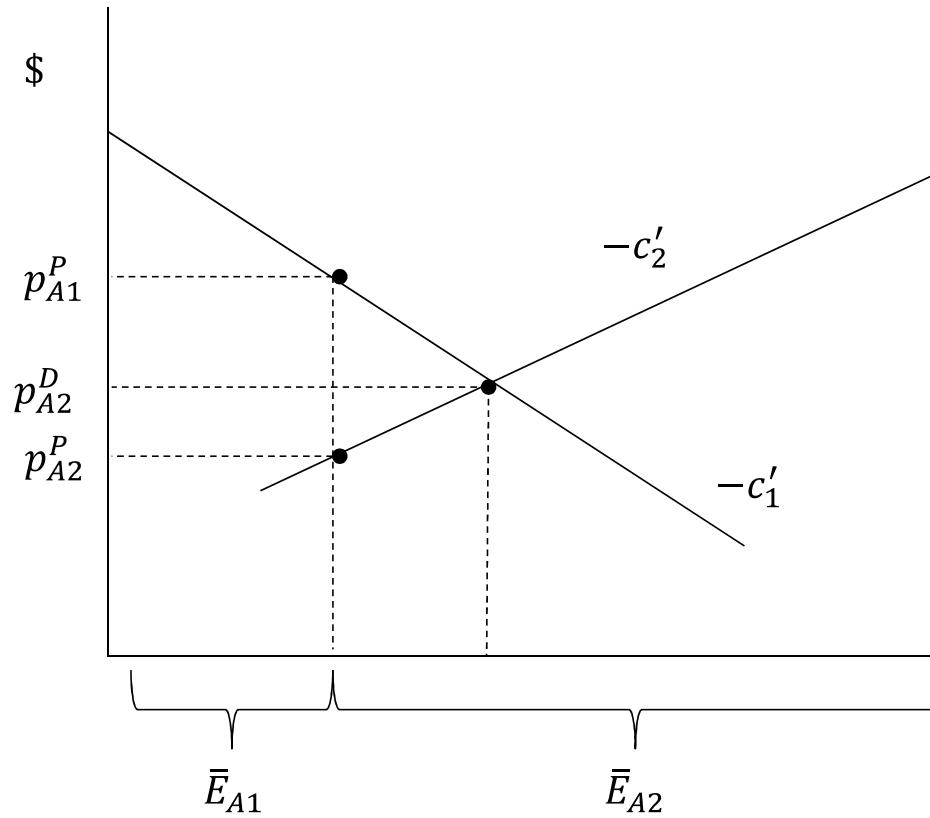
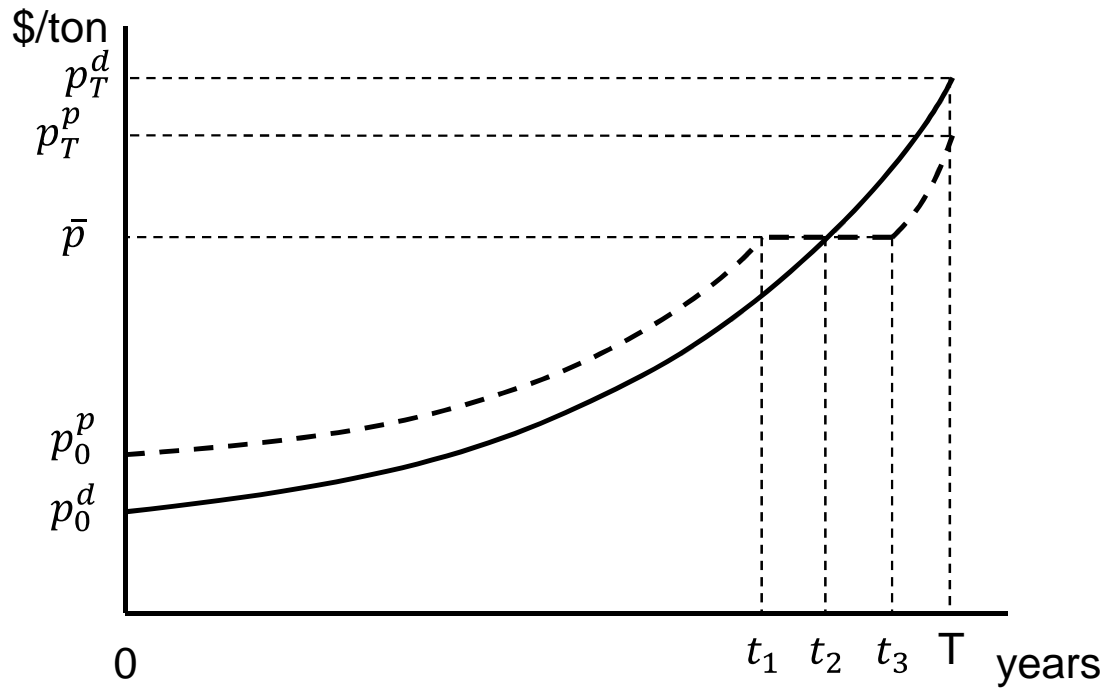


Figure 3: Equilibria for Prompt and Delayed Compliance with Delayed Allocations



Note: \bar{E}_{A1} is the amount of the cap allocated in period 1, and \bar{E}_{A2} is the amount of the cap allocated in period 2 (the delayed allocation). The superscript P indicates prompt compliance and the superscript D indicates delayed compliance.

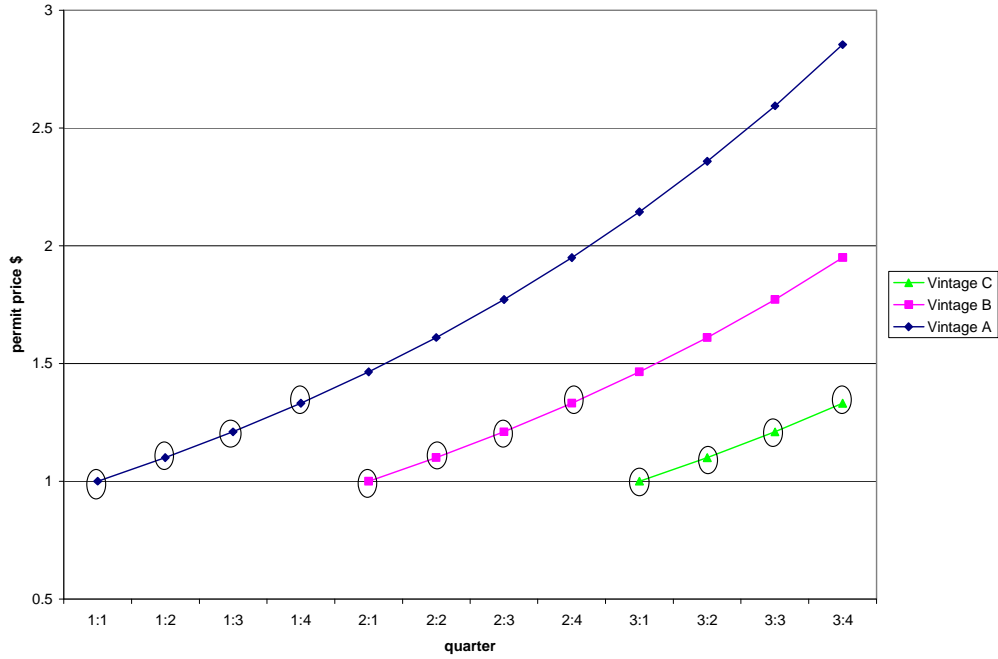
Figure 4: Price Paths with Price Ceiling \bar{p} for Prompt and Delayed Compliance



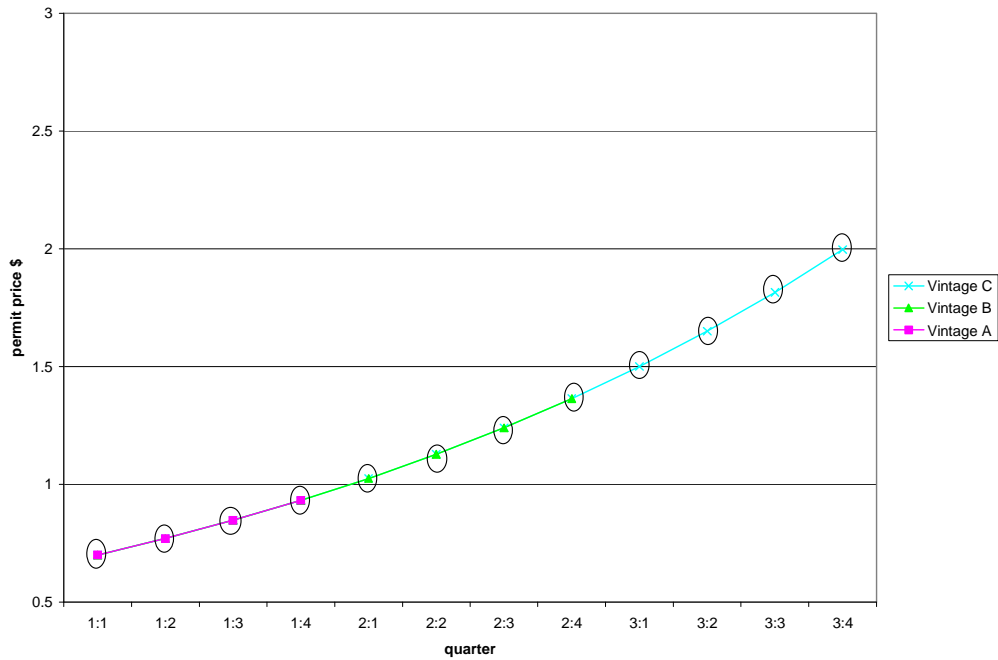
Note: The solid line illustrates delayed compliance. The dashed line illustrates prompt compliance with a small initial cap/large reserve fund.

Figure 5: Temporal Restrictions: Equilibrium Permit Price Paths for Four Market Designs

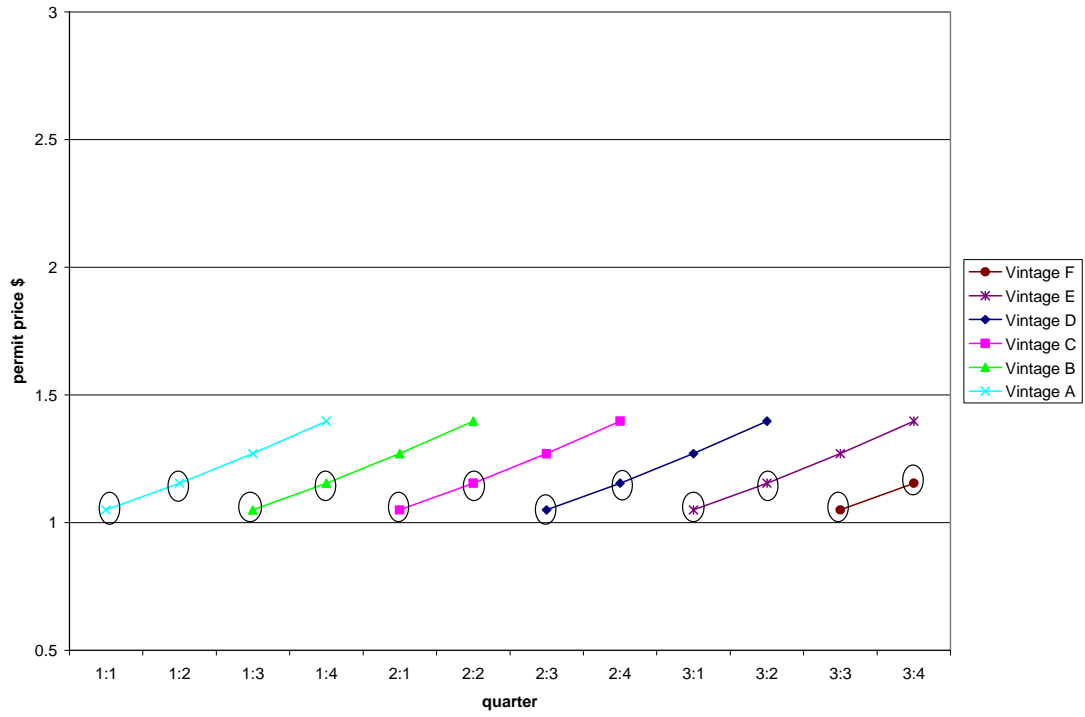
Panel A: Banking but no borrowing (ARP, CAIR, and CSAPR)



Panel B: Borrowing (with or without banking) (RGGI and AB 32)



Panel C: Overlapping permit vintages (RECLAIM)



Panel D: Costly borrowing with banking (Waxman-Markey)

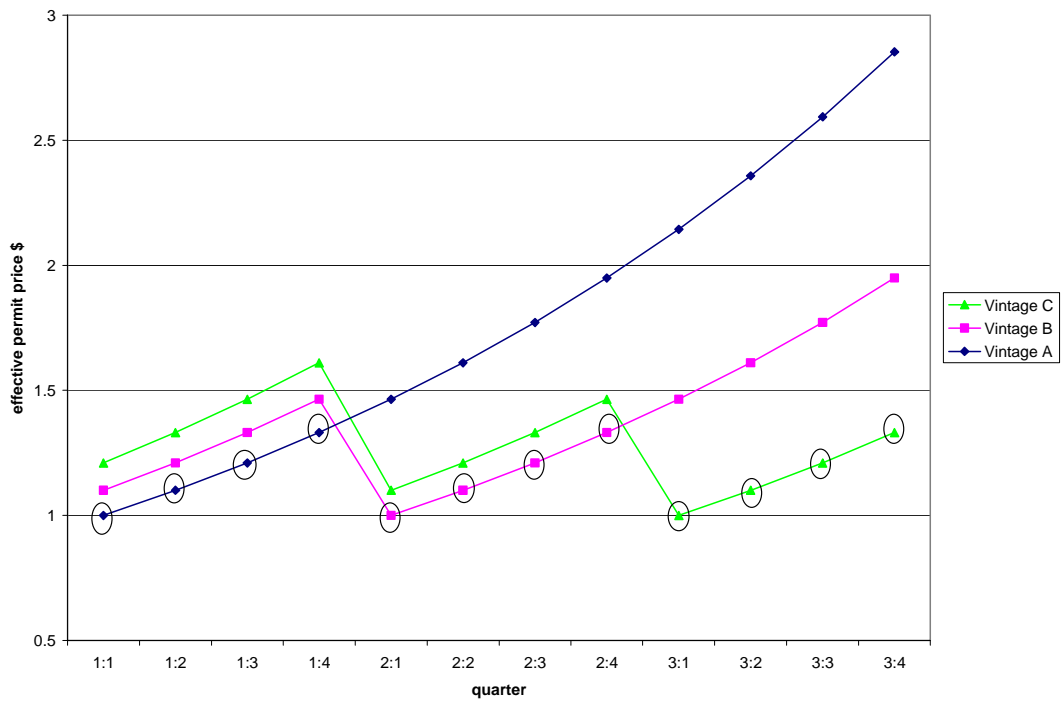


Figure 6: Spatial Restrictions: Marginal Abatement Costs and Marginal Damages for Two Heterogeneous Facilities

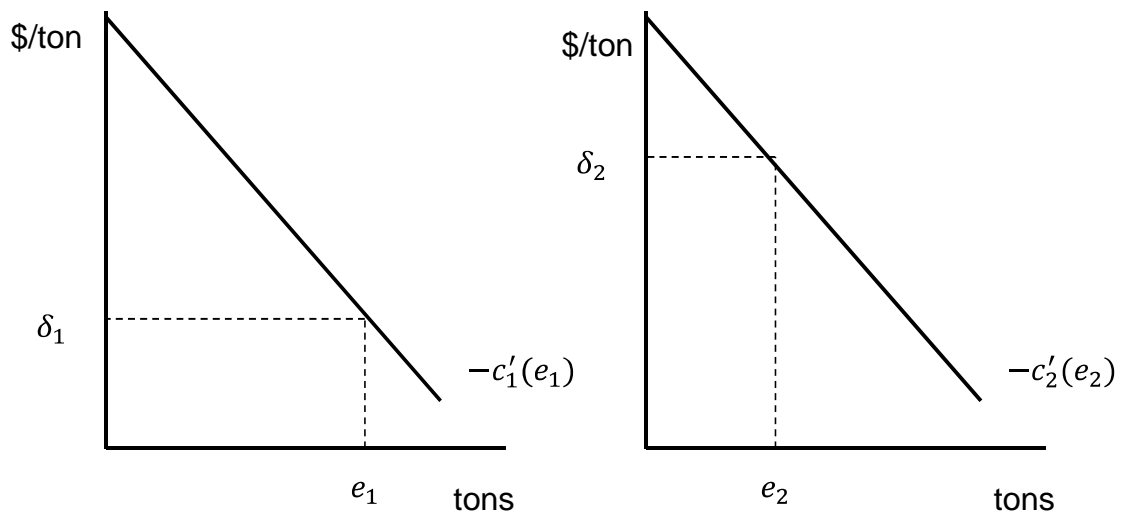


Table 1. Compliance Timing and Permit Validity in Cap-and-Trade Programs

Program (pollutant)	Compliance Timing		Permit Banking		Permit Borrowing		Spatial Limits Within Program
	Emissions Reporting ¹	Permit True Up	Explicitly Allowed? ²	Qualifications ³	Explicitly Allowed? ²	Qualifications ³	
Acid Rain Program (ARP) (sulfur dioxide)	quarterly	annual	yes	unlimited	no	none	no
NOx Budget Program (NBP) (nitrogen oxides)	quarterly	annual	yes	quantity tax on use of banked permits above a specified threshold	no	none	no
Clean Air Interstate Rule (CAIR) (nitrogen oxides and sulfur dioxide)	quarterly	annual	yes	unlimited	no	none	two NO _x markets in eastern U.S.
Cross-State Rule (CSAPR) (nitrogen oxides and sulfur dioxide)	quarterly	annual	yes	unlimited	no	none	two NO _x markets; two SO ₂ markets; variability limits on state emissions
RECLAIM (nitrogen oxides and sulfur dioxide)	quarterly	overlapping annual com- pliance cycles	no	limited ability to bank due to over- lapping permit cycles	no	limited ability to borrow due to over- lapping permit cycles	inland permits not valid in coastal zone
EU ETS (greenhouse gases)	annual	annual	yes	banking not allowed from first phase to second phase	no	unlimited borrowing from the next year's vintage of permits	no
Waxman-Markey (WM) (greenhouse gases)	quarterly	annual	yes	unlimited	yes	borrowing from the next year's vintage of permits; borrowing with interest from vintage years +2 to +5	no

RGGI (greenhouse gases)	quarterly	3-year period	yes	unlimited	no	unlimited borrowing within 3-year compliance period	no
California AB 32 (AB 32) (greenhouse gases)	annual	3-year period with 30% annual down payment	yes	unlimited	no	unlimited borrowing within 3-year compliance period conditional on annual down payment	includes elec- tricity imported to California

¹ This stage of program administration includes emissions reporting by regulated sources and emissions verification by the regulator.

² “Explicitly allowed?” asks whether the program allows banking or borrowing through an affirmative provision.

³ “Qualifications” describes explicit or implicit conditions on banking and borrowing.