Hotelling Under Pressure*

Soren T. Anderson† Ryan Kellogg‡ Stephen W. Salant§

November 27, 2016

Abstract

We show that oil production from existing wells in Texas does not respond to oil prices, while drilling activity and costs respond strongly. To explain these facts, we reformulate Hotelling's (1931) classic model of exhaustible resource extraction as a drilling problem: firms choose when to drill, but production from existing wells is constrained by reservoir pressure, which decays as oil is extracted. The model implies a modified Hotelling rule for drilling revenues net of costs, explains why the production constraint typically binds, and rationalizes regional production peaks and observed patterns of prices, drilling, and production following demand and supply shocks.

JEL classification numbers: Q3, Q4

Key words: crude oil prices; oil extraction; decline curve; oil drilling; rig rental rates; exhaustible resource

*For helpful comments and suggestions, we are grateful to Ying Fan, Cloé Garnache, Kenneth Hendricks, Stephen Holland, Lutz Kilian, Mar Reguant, Dick Vail, Jinhua Zhao, and participants at numerous seminars and conferences. We also thank the editor, Ali Hortaçsu, and three anonymous referees for many excellent suggestions. Outstanding research assistance was provided by Dana Beuschel, Dylan Brewer, and Sam Haltenhof. Part of this research was performed while Kellogg was at UC Berkeley's Energy Institute at Haas.

†Michigan State University and NBER. Email: sta@msu.edu

‡University of Chicago and NBER. Email: kelloggr@uchicago.edu

§University of Maryland, University of Michigan, and Resources for the Future. Email: ssalant@umich.edu

1 Introduction

Hotelling's (1931) classic model of exhaustible resource extraction, featuring forward-looking resource owners that maximize wealth by trading off extraction today versus extraction in the future, holds great conceptual appeal. Hotelling's logic has been applied to a wide range of resource extraction problems, including production from a crude oil reserve. Yet despite the theory's elegance, its mapping into empirical work in energy economics has been at best a limited success. The empirical literature on Hotelling has largely focused on testing the "Hotelling rule" that resource prices (or, more properly, in-situ values) should rise at the rate of interest, often finding that the rule fails to hold.[1] Meanwhile, a new micro-empirical energy economics literature focused on the oil and gas industry has flourished but has not linked itself to Hotelling's theory.[2] Our paper aims to bridge these two literatures by showing that a Hotelling model of oil drilling and production—suitably reformulated to reflect crucial features of petroleum geology and the industry's cost structure—generates a rich set of empirical predictions that are consistent with observables such as well-level production, drilling activity, drilling costs, and oil prices.

We begin by establishing two core empirical facts, using data on oil production and drilling in Texas from 1990–2007. First, we show that oil production from existing wells declines asymptotically toward zero and is almost completely unresponsive to oil price shocks.[3] This behavior is inconsistent with most Hotelling extraction models in the literature, which

[1] Barnett and Morse (1963), Smith (1979), Slade (1982), and Berck and Roberts (1996) find limited evidence for an upward trend in exhaustible resource prices, but tests based on prices alone are not correctly specified unless extraction costs are negligible. Papers estimating in-situ values, including Stollery (1983), Miller and Upton (1985), Farrow (1985), Halvorsen and Smith (1984), Halvorsen and Smith (1991), and Thompson (2001), find mixed results. See Krautkraemer (1998) and Slade and Thille (2009) for recent reviews.

[2] Kellogg (2011) and Covert (2015) examine learning and productivity in well drilling and fracking, respectively. Kellogg (2014) studies the effect of oil price volatility on drilling investments. Boomhower (2015) and Muehlenbachs (2015) study firms' decisions to either abandon or environmentally remediate wells that are no longer productive. Lewis (2015) studies the impacts of environmental restrictions on drilling in Wyoming. A number of very recent papers examine the effects of the shale boom on local economic and health outcomes. See, for example, Muehlenbachs, Spiller and Timmins (2015) and the citations contained therein.

[3] The one exception to this finding is that wells with very low productivity shut down during the oil price collapse of the late 1990s and later restarted when oil prices rebounded. We show that an extension to our baseline model that allows for well-level fixed operating costs can rationalize this behavior.

treat resource extraction as a "cake-eating" problem in which resource owners are able to allocate extraction across different periods without constraint—and we show that this behavior is not driven by common-pool problems, oil lease contract provisions, or other institutional factors. Second, we show that, in contrast to our results for oil production from existing wells, the drilling of new oil wells in Texas and the rental price of drilling rigs both respond strongly to oil price shocks.

We hypothesize that these empirical facts can be explained by several features of the oil extraction industry's structure that are well understood by petroleum geologists and engineers but that have received limited attention in the economics literature.[4] When a well is first drilled, the pressure in the underground oil reservoir is high. Production may therefore initially be rapid, since the maximum rate of fluid flow is roughly proportional to the pressure available to drive the oil through the reservoir, into the well, and then up to the surface. Over time, however, as extraction depletes reserves, the reservoir pressure declines, and the well's maximum flow decays toward zero. Oil extractors can rebuild their production capacity by drilling new wells. However, if the industry collectively drills at a faster rate, the per-well cost of drilling additional wells rises. Moreover, cumulative drilling is ultimately limited by the overall scarcity of oil.

We therefore recast Hotelling's (1931) canonical model, when applied to crude oil, as a problem in which resource owners can affect output through adjustments on two margins: well-level oil production (the intensive margin), and the rate of drilling new wells (the extensive margin). We retain Hotelling's conceptually appealing framework with forward-looking, wealth-maximizing agents constrained by finite resources. However, consistent with the industry's structure, we model a capacity constraint on oil extraction due to reservoir pressure that is proportional to the amount of recoverable oil remaining in existing wells.[5]

[4]Space constraints permit only a brief treatment here of the geologic and engineering basis for well-level capacity constraints. For a fuller discussion of fluid flow and production decline curves, Hyne (2001) is an excellent source that does not require a geology or engineering background.

[5]While appropriate for most crude oil extraction, our model is not applicable to resources such as coal, metal ores, or oil sands that are mined rather than produced through wells. For these resources, pressure-driven fluid flow is not important, and marginal extraction costs are likely to be substantial and increasing

In addition, we assume that the marginal cost of drilling a new well is strictly increasing in the aggregate rate of drilling, which is consistent with the upward-sloping supply curve for rig rentals implied by our data.

Using this model, we first characterize the incentive to produce from existing wells. The pressure constraint significantly dampens any incentive to withhold selling oil in the present to obtain a higher discounted price in the future, since deferred production can only be recovered gradually over a long time. Indeed, production at the constraint can be optimal even when prices rise faster than the rate of interest over an interval, provided the interval is sufficiently short. We tie this insight back to our oil production data by showing that, given observed oil prices during our sample, well owners never had an incentive to cut production below their declining capacity constraints, even during the late 1990s when oil prices temporarily were anticipated to rise much faster than the rate of interest. This finding is robust to several plausible assumptions for how oil producers historically may have formed beliefs about future oil prices. Our model therefore rationalizes our initially surprising empirical result that well-level oil production in Texas smoothly declines over time and does not respond to oil price shocks.

Thus, in practice, the intensive margin does not play a significant role in determining oil output. Instead, drilling incentives are crucial for determining production dynamics in oil markets, and oil extraction is more akin to a "keg-tapping" problem than a cake-eating problem: extractors choose when to drill their wells (or tap their kegs), but the flow from these wells is (like the libation from a keg) constrained due to pressure and decays toward zero as more oil is extracted. In the canonical Hotelling model, price net of marginal extraction cost must have the same present value whenever production occurs. In our reformulated model, we show that a modified Hotelling rule holds: the stream of revenues a well earns over its lifetime, net of the cost of drilling the well, must have the same present value whenever drilling occurs. This insight naturally leads to a prediction that aligns with our

with the production rate, so that modeling extraction requires a very different approach than that presented here.

second core empirical result: the rate of drilling and the rental cost of drilling rigs should respond positively to oil price shocks.

We also show that the equilibrium dynamics implied by our model easily and naturally replicate not only our two motivating empirical results, but also a wide range of other salient qualitative features of the crude oil extraction industry. First, since the flow constraint will typically bind in equilibrium, aggregate production will evolve gradually over time, following changes in the drilling rate, and will only respond to shocks with a significant lag. This result provides a foundation for a macro-empirical literature in energy economics showing that aggregate oil production is price inelastic, at least in the short run (Griffin 1985; Hogan 1989; Jones 1990; Dahl and Yucel 1991; Ramcharran 2002; Güntner forthcoming).[6] This inelasticity has important implications for the macroeconomic effects of oil supply and demand shocks, since inelastic supply and demand lead to volatile oil prices (see Hamilton 2009 and Kilian 2009). Second, within oil-producing regions, the model predicts the commonly observed phenomenon (Hamilton 2013) that production initially rises as drilling ramps up but then peaks and eventually declines as drilling slows down and the declining flow from existing wells dominates the production profile. Third, local supply shocks arising either from the discovery of new resources or from cost-reducing technical change lead in our model to a surge in drilling and then production, as recently happened with the U.S. shale boom. Fourth, we show that positive global demand shocks lead to an immediate increase in oil prices, drilling activity, and rig rental prices, and that oil prices may subsequently gradually fall if the increased rate of drilling causes production to gradually increase. These results are reversed for negative demand shocks, which can—if large enough—lead oil prices to rise faster than the rate of interest following the initial drop in price. These predicted responses to demand shocks match our data on drilling activity and rig rental prices, and are consistent with observed patterns in oil futures markets following large shocks.

Within the Hotelling literature, our model is most closely related to models in which

[6]Rao (2010), however, finds evidence using well-level data that firms can shift production across wells in response to well-specific taxes.

firms face convex investment costs to expand reserves and thereby reduce their marginal extraction costs (Pindyck 1978; Livernois and Uhler 1987; Holland 2008; Venables 2014), and to models in which production is directly constrained and firms face convex costs to expand production capacity (Gaudet 1983; Switzer and Salant 1986; Holland 2008).[7] Both types of models can generate initial periods of rising production and falling prices as reserves grow or as production capacity builds, followed by an inevitable decline (these models, like ours, also allow for periods of increasing production and declining prices following positive demand shocks). In these models, however, the eventual decline in production results from rising Hotelling scarcity rent rather than from a flow constraint that declines with cumulative extraction. Thus, in other models, whenever production is declining it must be responsive to demand shocks—a prediction inconsistent with our data. Moreover, the models in these papers do not admit the possibility of oil prices rising faster than the rate of interest in equilibrium—a phenomenon implied by our futures price data. Finally, our model links capacity expansion to drilling activity and the marginal cost of capacity expansion to the rental rates on drilling rigs, each of which can be observed empirically.

Other papers in the economics literature, like ours, draw from the petroleum geology and engineering concept of a production decline curve. Nystad (1987), Adelman (1990), Black and LaFrance (1998), Davis and Cairns (1998), Cairns and Davis (2001), Thompson (2001), Gao, Hartley and Sickles (2009), Smith (2012), Mason and van't Veld (2013), Cairns (2014), Ghandi and Lin (2014), and Okullo, Reynès and Hofkes (2015) are all premised on an oil production constraint that decays with cumulative extraction.[8] Our paper differs from this literature in several ways. First, we directly link our model to observables on oil production, drilling, and drilling costs. Our core empirical findings on the price responsiveness of well-level oil production, drilling activity, and drilling costs have not, to the best of our knowledge,

[7]For a review of the theoretical Hotelling literature through the 1970s, see Devarajan and Fisher (1981). For more recent reviews, see Krautkraemer (1998), Gaudet (2007), and Gaudet and Salant (2014). For an accessible theoretical primer, see Salant (1995).

[8]In addition, Venables (2014) studies exponential decline from existing wells as a special case of its model, in which ultimate recovery depends on the rate of resource depletion. Prices are endogenous, but drilling costs are exogenous.

been documented in prior work.[9] Second, motivated by our empirical results, we develop a reformulated Hotelling model that emphasizes the rate of drilling as the central choice variable rather than the rate of oil production, and we use our model to explain why the production constraint binds empirically, even during periods in our sample when the oil price is plausibly anticipated to rise faster than the rate of interest. Third, we show that Hotelling's logic still applies in our model, with the discounted marginal revenue stream from drilling minus the marginal cost of drilling (i.e., the rig rental price) rising at the interest rate. Finally, we show that the drilling, production, and price dynamics implied by our model can match those of the real-world oil extraction industry, including responses to demand and supply shocks.[10]

Overall, our paper demonstrates that a Hotelling-style model that is grounded in the actual cost structure of the oil industry can deliver a diverse set of implications that are borne out in industry data. The model we study remains a simplification in many ways that are no doubt important; for instance, it abstracts away from reserve heterogeneity, uncertainty about future demand and supply, and capital investment in drilling rigs. Our hope is that the results presented here will renew interest in using Hotelling models—and in extending ours—to better understand and predict the behavior of oil extractors and markets.

[9]Thompson (2001) and Mason and van't Veld (2013) provide related evidence that the ratio of aggregate production to proven reserves has remained roughly constant over time, even in the face of large price variation. Thompson (2001) also presents results that a decline curve model outperforms competing models in explaining cross-sectional variation in the market values of oil-producing firms, and shows that natural gas consumption and prices are seasonal, while production is not. Black and LaFrance (1998) use a structural model to test a null hypothesis that production from fields in Montana can be explained by decline curves against the alternative that oil prices matter. They reject the null, perhaps because production from new wells and re-entries responds to prices, or perhaps due to the functional form assumptions that are imposed.

[10]Of the sources cited above, Okullo et al. (2015) and Mason and van't Veld (2013) have models that are most similar to ours and consider equilibrium outcomes. But Okullo et al. (2015) focus on long-run dynamics and do not model the response to shocks, while Mason and van't Veld (2013) only derive equilibrium outcomes in a simplified, two-period version of their model. In addition, both papers permit non-trivial marginal production costs below the constraint, which is a case rejected by our data.

2 Empirical evidence from Texas

In this section, we document our fundamental empirical results that oil production exhibits nearly zero response to oil price shocks, but drilling activity—along with the cost of renting drilling rigs—responds strongly. We then discuss intuitively how these results derive from the technology of crude oil extraction before turning to a formal model in section 3.

2.1 Data sources

Our crude oil drilling data for 1990–2007 come from the Texas Railroad Commission's (TRRC's) "Drilling Permit Master" dataset, which provides the date, county, and lease name for every well drilled in Texas. A lease is land upon which an oil production company has obtained the right to drill for and produce oil and gas. Over 1990–2007, 157,270 new wells were drilled,[11] and there were also 40,760 "re-entries" of existing wells.[12] Oil production data come from the TRRC's "Oil and Gas Annuals" dataset, which records monthly crude oil production at the lease level.[13] Individual wells are not flow-metered, so we generally cannot observe well-level production.[14]

Our analysis of oil production focuses on whether firms respond to oil price shocks by adjusting the flow rates of their previously drilled wells, possibly all the way to zero, which is known as "shutting in" a well.[15] Reducing a well's flow would typically be accomplished by

[11] A small share ($< 10\%$) of the new wells were drilled to inject water or gas into the reservoir rather than extract oil. These injection investments can mitigate, but not eliminate, the rate of production decline that we document here.

[12] A re-entry occurs when a rig is used to deepen a well, drill a "sidetrack" off an existing well bore, stimulate production by fracturing the oil reservoir, or otherwise re-complete the well.

[13] Due to false zeros for some leases in 1996 and December 2004–2007, we augmented these data by scraping information from the TRRC's online production query tool, verifying that the two sources match for leases and months not affected by the data error.

[14] Direct production includes oil, gas, and often water. Separation of these products typically occurs at a single facility serving the entire lease, with the oil flowing from the separation facility into storage tanks. Oil is metered leaving the storage tank for delivery to a pipeline or tanker truck for sale.

[15] Throughout our analysis, we assume that oil price movements are exogenous to actions undertaken by Texas oil producers. This treatment is plausible given that Texas firms are a small share of the world oil market (1.3% in 2007) and evidence that oil price shocks during our sample were primarily driven by global demand shocks and (to a lesser extent) international rather than U.S. supply shocks (Kilian 2009). Moreover, the positive covariance between drilling activity and oil prices apparent in figure 2 strongly suggests that

choking off the well (if the well was flowing naturally) or by slowing down the pumping unit (if it was being pumped).[16] To distinguish these actions from investments in new production, such as drilling a new well, we focus our analysis on leases in which no rig work took place,[17] including re-entries.[18] Of the 33,108 leases in the production data for which production volumes are not missing for any month from 1990–2007 and production is non-zero for at least one month, 16,159 leases (48.8%) did not experience rig work from 1990–2007. The average daily production rate for these leases is 3.6 barrels of oil per day (bbl/d), with a standard deviation of 18.2 bbl/d.[19] This low average production rate reflects the fact that most oil fields in Texas are mature and were discovered long ago. We find that 1,071,229 (31%) of the observed lease-months have zero production, while the maximum is 9,510 bbl/d.

Our oil price data come from the New York Mercantile Exchange (NYMEX) prices for West Texas Intermediate (WTI) crude oil, covering 1990–2007. We use the front-month (upcoming month) futures price as our measure of the spot price of crude oil,[20] and we use the All Urban, All Goods Less Energy Consumer Price Index (CPI) of the Bureau of Labor Statistics to convert all prices to December 2007 dollars. Figure 1 shows that crude oil prices varied considerably over the 1990–2007 sample. The price of oil fell substantially during 1998–1999, which Kilian (2009) attributes to a negative demand shock arising from the

Texas drilling activity is responding to price shocks rather than vice versa.

[16]The vast majority of the wells in the dataset are pumped. The average lease-month in the data has 2.02 pumped wells and 0.06 naturally flowing wells.

[17]Identifying leases with rig work requires matching the drilling dataset to the production dataset. Since lease names are not consistent across the two datasets, we conservatively identify all county-firm pairs in which rig work took place and then discard *all* leases corresponding to such pairs (unlike leases, counties and firms are consistently identified with numeric codes in both datasets).

[18]We exclude leases with re-entries because re-entries are economically similar to the drilling of a new well: they are an investment that increases the amount of productive capacity, enabling a larger stream of future oil production. The cost of a re-entry is smaller than that of drilling a new well, but it is still an investment rather than an operational change in production rate. That is, re-entries do not allow oil production to respond immediately to oil price shocks. Rather, like new drilling investments, re-entries allow production to build up gradually over time as new capacity is added.

[19]Excluded leases (i.e., those with rig work) average 6.3 bbl/d. In appendix A, we examine subsets of the leases in our analysis dataset that have relatively high production rates and are therefore more similar to the excluded leases. We obtain similar results to what we discuss in section 2.2.

[20]In section 4 below, in which we test our model's predictions for firms' production decisions, we will also make use of prices for longer-term futures contracts to derive measures of firms' future price expectations, in some cases accounting for non-diversifiable risk in oil markets.

Asian financial crisis. Conversely, oil prices rapidly increased during the mid-2000s; Kilian (2009) and Kilian and Hicks (2013) attribute this increase to a series of large, positive, and unanticipated shocks to the demand for oil, primarily from emerging Asian markets.

Finally, we obtained information on rental prices ("dayrates") for drilling rigs from RigData (1990–2013). As discussed in Kellogg (2011), the oil production companies that make drilling and production decisions do not drill their own wells but rather contract drilling out to independent service companies that own rigs. Rental of a rig and its crew is typically the largest line item in the overall cost of a well. The RigData data are quarterly, covering Q4 1990 through Q4 2007.[21] Observed dayrates range from $6,113 to $15,168 per day, with an average of $8,903 (all real December 2007 dollars).

2.2 Production from existing wells does not respond to prices

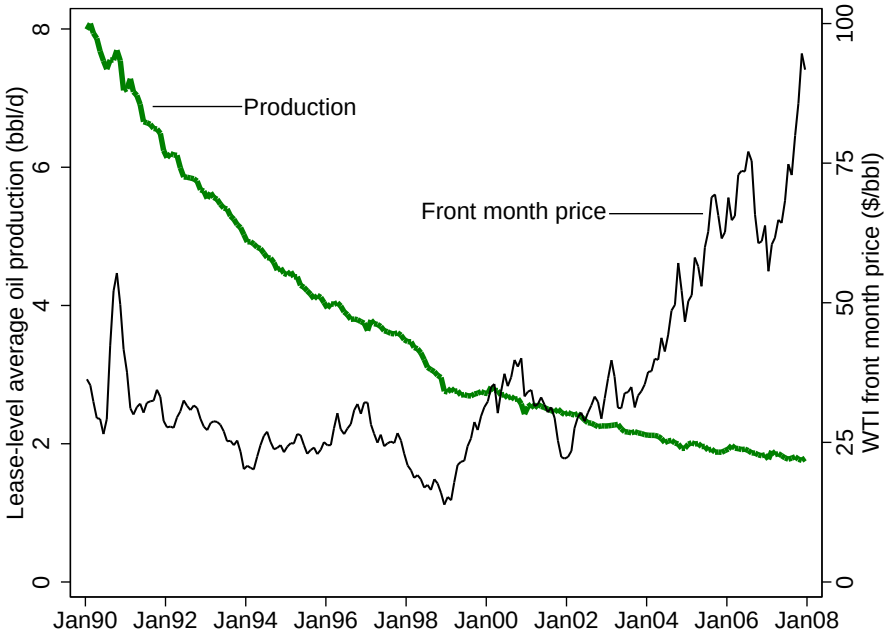
Our main empirical results focus on production from leases on which there was no rig activity from 1990–2007, so that all production comes from pre-existing wells. Figure 1 presents daily average production (in bbl/d) for these leases in each month, along with monthly crude oil front month ("spot") prices. Production is dominated by a long-run downward trend, with little response to the spot price of oil. In appendix A, we present regression results confirming the lack of response to price incentives.[22] We also demonstrate that the pattern shown in figure 1 holds for subsamples of leases that have relatively high production volumes and for production from wells that are drilled during the sample period. Thus, our results are not specific only to the low-volume wells that predominate in Texas.

Figure 1 does suggest that the production decline rate may have accelerated briefly during the 1998–1999 period in which the spot price fell below $20/bbl. In appendix C, we study this period further and find that this production dip is driven by shut-ins of low volume,

[21] The RigData data are broken out by region and rig depth rating. We use dayrates for rigs with depth ratings between 6,000 and 9,999 feet (the average well depth in our drilling data is 7,424 feet) for the Gulf Coast / South Texas region.

[22] Our regressions also show that production does not respond to anticipated changes in future oil prices as reflected in longer-term futures prices.

Figure 1: Crude oil prices and production from existing wells in Texas



Note: This figure presents crude oil front month ("spot") prices and daily average lease-level production from leases on which there was no rig activity (so that all production comes from pre-existing wells). All prices are real $2007. See text for details.

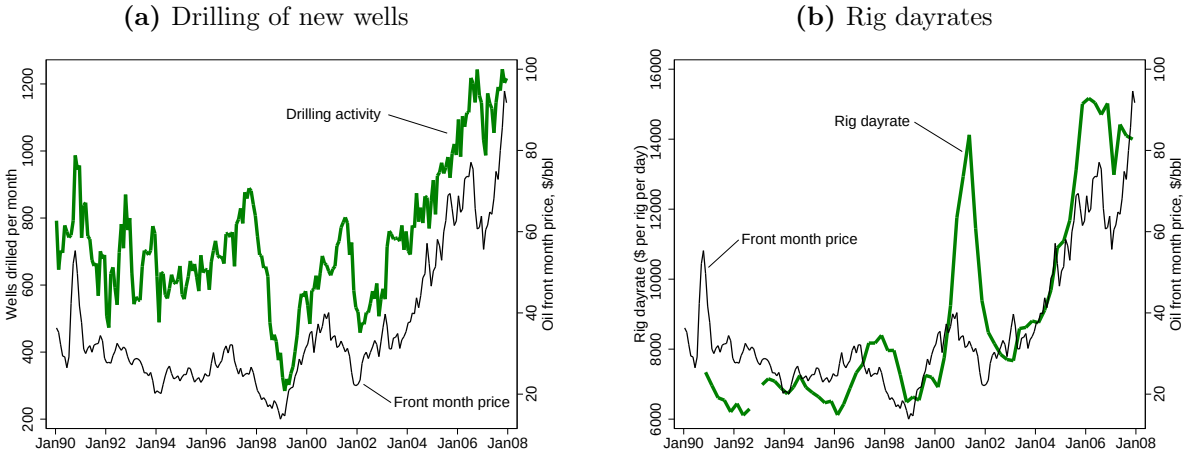
marginal wells.

2.3 Rig activity does respond to price incentives

These no-response results for production from existing wells contrast starkly with the price-responsiveness of drilling activity in Texas. Figure 2(a) shows a pronounced positive correlation between the spot price of crude oil and the number of new wells drilled each month across all leases in our data. Appendix A presents related regression results indicating that the elasticity of the monthly drilling rate with respect to the crude oil spot price is about 0.7 and statistically different from zero. We have also found that the use of rigs to re-enter old wells correlates with oil prices, though not as strongly as the drilling of new wells (see figure 9 in appendix A).

When oil production companies drill more wells in response to an increase in oil prices,

Figure 2: Texas rig activity versus crude oil spot prices



Note: Panel (a) shows the total number of new wells drilled across all leases in our dataset. Panel (b) shows dayrates for the Gulf Coast / South Texas region, for rigs with depth ratings between 6,000 and 9,999 feet. The dayrate data are quarterly rather than monthly. Data are available beginning in Q4 1990, and data for Q4 1992 are missing. Oil prices are real $2007. See text for details.

more rigs (and crews) must be put into service to drill them. Figure 2(b) shows that these fluctuations in rig demand are reflected in a positive covariance between rig dayrates and oil prices. Regressions confirm that the elasticity of the rig rental rate with respect to oil prices is large (0.77) and statistically significant.

2.4 Industry cost structure explains these price responses

The analysis above documents that: (1) production from drilled wells is almost completely unresponsive to changes in the front-month oil price, with an exception being an increased rate of shut-ins during the 1998 oil price crash; and (2) drilling of new wells responds strongly to oil price changes, and rig dayrates respond commensurately. Here, we argue that these empirical results reflect an industry cost structure with the following characteristics:[23]

1. The rate of production from a well is physically constrained, and this constraint declines

[23]For a particularly cogent discussion within the economics literature, see Thompson (2001).

asymptotically toward zero as a function of cumulative production. This function is known in the engineering literature (Hyne 2001) as a well's production decline curve.

2. The marginal cost of production below a given well's capacity constraint, consisting mainly of energy input to the pump (if there is one) and the cost of transporting oil from the lease to oil purchasers, is very small relative to observed oil prices.

3. The fixed costs of operating a producing well are non-zero. There may also be costs for restarting a shut-in well, but they are not too large to be overcome. We return to this issue in appendix C.

4. Drilling rigs and crews are a relatively fixed resource, at least in the short run. Higher rental prices are required to attract more rigs into active use, leading to an upward-sloping supply curve of drilling rigs for rent.

The capacity constraint and low marginal production cost relate to the observation that oil production from existing wells steadily declines and does not respond to price shocks, contradicting predictions from standard Hotelling models with increasing marginal extraction costs. Because oil producers in Texas are price-takers,[24] production will be unresponsive to price shocks, as the data reflect, only if the oil price intersects marginal cost at a vertical, capacity-constrained section of the curve. While the marginal cost of production below the capacity constraint is not necessarily zero, it must be well below the range of oil prices observed in the data.[25]

[24]The market for crude oil is global, and Texas as a whole (let alone a single firm) constitutes only 1.3% of world oil production; thus, the exercise of market power by Texas oil producers is implausible (Texas and world production data are for 2007 and were accessed from the U.S. Energy Information Administration at `http://www.eia.gov/dnav/pet/pet_crd_crpdn_adc_mbblpd_a.htm` and `http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=50&pid=53&aid=1`, respectively, on 27 September, 2015).

[25]For oil reservoirs produced with the help of injection wells, the injection rate may be sensitive to the oil price if there are high marginal costs of injection. This injection rate price sensitivity would result in oil production price sensitivity. The lack of price response in our data suggests that this issue is not important overall for Texas production, but it may be important for particular enhanced oil recovery projects, in Texas or elsewhere.

The existence of a capacity constraint for well-level production is consistent not only with the data presented above but also with standard petroleum geology and engineering. As noted recently in the economics literature by Mason and van't Veld (2013), the flow of fluid through reservoir rock to the well bore is governed by Darcy's law (Darcy 1856), which stipulates that the rate of flow is proportional to the pressure differential between the reservoir and the well.[26] In the simplest model of reservoir flow, the reservoir pressure is proportional to the volume of fluid in the reservoir. In this case,[27] the maximum flow rate is proportional to the remaining reserves, consistent with an exponential production decline curve and with the stylized fact reported in Thompson (2001) and Mason and van't Veld (2013) that U.S. production has remained close to 10% of proven reserves since the industry's infancy, despite large changes in production over time.[28]

As we show in appendix C, some relatively low-volume wells were shut in during 1998. These shut-ins are consistent with the existence of fixed production costs, which intuitively arise from the need to monitor and maintain surface facilities so long as production is non-zero. When the oil price fell in 1998, production from these wells may no longer have been sufficient to cover their fixed costs, rationalizing their shut-in. When oil prices subsequently recovered, many (though not all) of these wells restarted, suggesting that start-up costs can sometimes be overcome.

In appendix A, we consider and rule out alternative explanations for the lack of response of oil production to oil prices. We show that our results cannot be explained by (1) leasing agreements that require non-zero production (because multiple-well leases show the same results); (2) races-to-oil induced by open-access externalities within oil fields (because fields

[26]Darcy's law governs oil flowing through the reservoir and into the bottom of the well bore. Installing a pump on a well effectively eliminates the need for the oil to overcome gravity as it rises up the well, but it does not negate Darcy's law, implying that one could never drain all of the oil in finite time, even if the pump pulled a vacuum on the bottom of the well. Enhanced oil recovery through the use of injection wells can slow the pressure decline in the reservoir, but again oil production must inevitably decline as the production wells produce more and more of the injected fluids rather than oil.

[27]More complex cases, which might involve the presence of gas, water, or fractures in the reservoir, may yield a more general hyperbolic decline.

[28]Thompson (2001) shows that daily production in the 1980s and 1990s hovered near 0.03% of reserves, which implies annual production equal to $365 \cdot 0.03\% = 10.9\%$ of annual capacity.

controlled by a single operator show the same results); or (3) well-specific production quotas (because production quotas are far from binding).

3 Recasting Hotelling as a drilling problem

In this section, we develop a theory of optimal oil drilling and extraction that closely follows the industry cost structure described above. After formulating the problem, we derive and interpret conditions that necessarily hold at any optimum, focusing first on incentives to produce at the capacity constraint and then on incentives to drill new wells.

3.1 Planner's problem and necessary conditions

Because there are millions of operating oil wells in the world, we formulate our model as a decision problem in which there is a continuum of infinitesimally small wells to be drilled. We use continuous time to facilitate interpretation of the necessary conditions and the analysis of equilibrium dynamics. The planner's problem is given by:

$$\max_{F(t),a(t)} \int_{t=0}^{\infty} e^{-rt} \left[U(F(t)) - D(a(t)) \right] dt \tag{1}$$

subject to

$$0 \leq F(t) \leq K(t) \tag{2}$$

$$a(t) \geq 0,\ R(t) \geq 0 \tag{3}$$

$$\dot{R}(t) = -a(t),\ R_0 \text{ given} \tag{4}$$

$$\dot{K}(t) = a(t)X - \lambda F(t),\ K_0 \text{ given}, \tag{5}$$

where $F(t)$ is the rate of oil flow at time t (a choice variable), $a(t)$ is the rate at which new wells are drilled (a choice variable), $K(t)$ is the capacity constraint on oil flow (a state

variable), and $R(t)$ is the measure of wells that remain untapped (a state variable). The instantaneous utility derived from oil flow is given by $U(F(t))$, where $U(\cdot)$ is strictly increasing and weakly concave; we normalize $U(0) = 0$. The total instantaneous cost of drilling wells at rate $a(t)$ is given by $D(a(t))$, where $D(\cdot)$ is strictly increasing and weakly convex, and $D(0) = 0$. We denote the derivative of the total drilling cost function as $d(a(t))$ and assume that $d(0) \geq 0$. Utility and drilling costs are discounted at rate r.

Consistent with our empirical results from Texas, we assume a trivially low (i.e., zero) marginal cost of extraction up to the constraint.[29] We provisionally ignore any fixed costs for operating, shutting in, or restarting wells because such costs are only relevant for marginally productive wells or when oil prices are very low.[30] We show in appendix C that fixed costs therefore do not have a qualitatively important impact on drilling incentives, since newly drilled wells will typically only become marginal many years after drilling.

Condition (4) describes how the stock of untapped wells $R(t)$ evolves over time. The planning period begins with a continuum of untapped wells of measure R_0, and the stock thereafter declines one-for-one with the rate of drilling. Condition (5) describes how the oil flow capacity constraint $K(t)$ evolves over time. The planning period begins with a capacity constraint K_0 inherited from previously tapped wells. The maximum rate of oil flow from a tapped well depends on the pressure in the well and is proportional, with factor λ, to the oil that remains underground. Thus, oil flow $F(t)$ erodes capacity at rate $\lambda F(t)$.[31] The planner can, however, rebuild capacity by drilling new wells. The rate of drilling $a(t)$ relaxes the capacity constraint at rate X, where we interpret X as the maximum flow from a newly

[29] As indicated above, marginal extraction costs are not literally zero. We ignore per-barrel extraction costs from existing wells because the lack of response to oil prices for such wells implies that marginal costs are low relative to oil prices.

[30] Accounting for these costs would complicate the analysis substantially. We would need to model, at each t, how the quantity of oil reserves remaining in tapped wells is distributed across the continuum of tapped wells, along with the shadow opportunity cost associated with extracting more oil from every point in this distribution.

[31] We assume proportional decay (i.e., an exponential production decline curve for wells producing at capacity) because doing so implies that aggregate capacity across all wells is a sufficient state variable for our problem. If we assumed more general forms of decay (e.g., a hyperbolic production decline curve), we would need to model the distribution of capacities across wells (a potentially infinite-dimensional space).

drilled well (or to be more precise, a unit mass of newly drilled wells).[32] If no new wells are being drilled at t ($a(t) = 0$) and production is set at the constraint ($F(t) = K(t)$), then oil flow decays exponentially toward zero at rate λ. If instead production is set to zero for an interval, then in the absence of drilling, reserve depletion ceases and hence the maximum flow does not change during that interval.

The total amount of oil in untapped wells is given by $R(t)X/\lambda$, so that the total amount of oil underground at the outset of the planning period is given by $Q_0 = (K_0 + R_0X)/\lambda$.[33] Because the flow capacity constraint is proportional to the remaining reserves, the total underground stock of oil will never be exhausted in finite time.

We assume that there is no above-ground storage of oil to focus our analysis and discussion on the implications of our model for extraction and drilling dynamics. Extending the model to include above-ground storage with an iceberg storage cost is straightforward, and we do so in appendix E.[34]

The solution to our planner's problem can, via the First Welfare Theorem, also be interpreted as the competitive equilibrium that would arise in a decentralized problem with continua of infinitesimally small consumers and private well owners (and no common pool problems), each of whom discounts utility or profit flows at the rate r. In a market context, consumers have an inverse demand function $P(F)$, well owners maximize their wealth by choosing feasible drilling and production paths, and heterogeneous rig owners rent out their drilling rigs, with each of these agents taking as given the time paths of the oil price and

[32] If the drilling cost function $D(a)$ is strictly convex, the planner would never find it optimal to set up a mass of wells instantaneously at $t = 0$ or at any other time, and the stock of untapped wells and oil flow capacity constraint would both evolve continuously over time. When the drilling cost is linear, however, such "pulsing" behavior may be optimal, leading to discontinuous changes in these state variables.

[33] A mathematically equivalent formulation of our problem would involve imposing resource scarcity directly on the recoverable oil stock remaining by replacing condition (4) with $\dot{Q}(t) = -F(t)$, where $Q(t) = (K(t) + R(t)X)/\lambda$ is the total amount of oil remaining underground at time t. We find that our current formulation leads to necessary conditions that are easier to interpret and manipulate.

[34] Intuitively, the presence of costly above-ground storage places an upper bound on the rate at which the oil price can increase in equilibrium. Appendix E demonstrates that our result that constrained production can be optimal even when the oil price is rising faster than the rate of interest (over a finite time interval) still holds when costly above-ground storage is available.

the rig rental rate. In equilibrium, the markets for crude oil and for rig rentals clear.[35] The equilibrium oil price may be inferred from the planner's marginal utility $U'(F(t))$, and the equilibrium rig rental rate may be inferred from the planner's marginal drilling cost $d(a(t))$.[36] We abstract away from any investment decisions of the heterogeneous rig owners and simply assume that they rent out their scarce equipment as long as the marginal cost of supplying it does not exceed the rig rental rate.[37] In the discussion below, we will primarily use the language of the planner's maximization problem, though we will find it convenient to use the competitive equilibrium language (and the notation $P(F)$ rather than $U'(F)$) when we discuss well owners' production incentives, taking the oil price path as given.

Following Léonard and Long (1992), the current-value Hamiltonian-Lagrangean of the planner's maximization problem is given by:

$$H = U(F(t)) - D(a(t)) + \theta(t)[a(t)X - \lambda F(t)] + \gamma(t)[-a(t)] + \phi(t)[K(t) - F(t)], \qquad (6)$$

where $\theta(t)$ and $\gamma(t)$ are the co-state variables on the two state variables $K(t)$ and $R(t)$, and $\phi(t)$ is the shadow cost of the oil flow capacity constraint.

Necessary conditions are given by equations (7) through (14) and are interpreted in

[35]The price-taking assumption on both sides of the market is reasonable for onshore Texas, given the existence of thousands of oil producing firms. In other areas, such as the deepwater Gulf of Mexico, only very large "major" firms participate, and these firms may be able to exert monopsony power in the rig market even if they are oil price takers. Finally, large OPEC nations such as Saudi Arabia can potentially exert market power in the global oil market.

[36]There also exist non-rig costs associated with drilling, such as materials and engineering costs. Thus, $d(a(t))$ can be viewed as the sum of these costs, which are invariant to $a(t)$, with the drilling rig rental cost.

[37]A richer model would allow for investment in durable drilling rigs; we save this extension for future work.

sections 3.2 and 3.3 below:

$$F(t) \geq 0,\ U'(F(t)) - \lambda\theta(t) - \phi(t) \leq 0,\ \text{comp. slackness } (c.s.) \tag{7}$$

$$K(t) - F(t) \geq 0,\ \phi(t) \geq 0,\ c.s. \tag{8}$$

$$a(t) \geq 0,\ \theta(t)X - d(a(t)) - \gamma(t) \leq 0, c.s. \tag{9}$$

$$\dot{R}(t) = -a(t),\ R_0 \text{ given} \tag{10}$$

$$\dot{\gamma}(t) = r\gamma(t) \tag{11}$$

$$\dot{K}(t) = a(t)X - \lambda F(t),\ K_0 \text{ given} \tag{12}$$

$$\dot{\theta}(t) = -\phi(t) + r\theta(t) \tag{13}$$

$$K(t)\theta(t)e^{-rt} \to 0 \text{ and } R(t)\gamma(t)e^{-rt} \to 0 \text{ as } t \to \infty. \tag{14}$$

The solution to these conditions will be unique under weak sufficient conditions.[38]

3.2 Implications of necessary conditions for production

We begin by focusing on condition (7), which characterizes production incentives. This condition involves the co-state variable $\theta(t)$, which denotes the marginal value of an addition to capacity at time t. This marginal value equals the additional stream of discounted future utility that can be obtained by producing oil optimally given the additional capacity. If the optimal program calls for strictly positive production at all times $\tau \geq t$, then $\theta(t)$ is simply given by the value of the stream of future marginal utilities $U'(F(\tau))$ discounted at the rate $r + \lambda$. If, on the other hand, it is optimal to shut in production for some interval, then this

[38] In particular, the solution is unique if: (1) U' and $-d$ are both strictly decreasing, since then the Hamiltonian-Lagrangian is strictly concave in the choice variables (Léonard and Long 1992); (2) U' and $-d$ are both globally constant and $U'X/(r+\lambda) \neq d$, in which case drilled wells are always produced at capacity and the initial stock of undrilled wells is either drilled immediately (if $U'X/(r+\lambda) > d$) or left undrilled forever (if $U'X/(r+\lambda) < d$) (in the knife-edge case that $U'X/(r+\lambda) = d$, any pattern of drilling combined with maximal production from drilled wells is optimal, so the solution is not unique); (3) U' is globally constant and $-d$ is strictly decreasing (see section 5); or (4) $-d$ is globally constant and U' is strictly decreasing (see appendix D). Case (1) is the empirically relevant case when our model is applied to global oil markets, while case (3) is the empirically relevant case when our model is applied to a local, oil-producing region.

stream of discounted marginal utilities serves as a lower bound on $\theta(t)$. Thus, we have:[39]

$$\theta(t) \geq \int_t^{\infty} U'(F(\tau))e^{-(r+\lambda)(\tau-t)}d\tau, \text{ holding with equality if } F(\tau) > 0 \text{ for all } \tau \geq t. \quad (15)$$

Intuitively, the marginal unit of capacity can always be used continuously from date t onward, generating wealth equal to the right-hand side of equation (15). If it is instead optimal to shut in and thereby defer production to some future interval, doing so must generate even greater wealth, yielding the inequality in (15).

This understanding of $\theta(t)$ facilitates the interpretation of condition (7). Increasing production at time t reduces the underground pressure and hence tightens the constraint on future oil flow at rate λ. Thus, the product $\lambda\theta(t)$ captures the opportunity cost of a marginal increase in flow at t in terms of forgone future utility. This marginal cost is independent of the rate of current production, while the marginal benefit decreases in $F(t)$. It then follows that if $U'(F(t)) - \lambda\theta(t) > 0$ for $F(t) = K(t)$, optimal production occurs at the upper bound $K(t)$ (i.e., production is capacity constrained at the optimum). Alternatively, if $U'(F(t)) - \lambda\theta(t) < 0$ for $F(t) = 0$, then optimal production is at the lower bound of zero. Finally, an interior solution is permitted if $U'(F(t)) - \lambda\theta(t) = 0$ for some $F(t) \in [0, K(t)]$. Condition (7) covers all three of these cases at once because (8) implies that $\phi(t) \geq 0$ is strictly positive only if $F(t) = K(t)$.

If $F(t) < K(t)$, then condition (13) implies that $\dot{\theta}(t)/\theta(t) = r$. If, in addition, $F(t) > 0$ then condition (7) implies that $U'(F(t))$ rises in percentage terms at the rate of interest.

[39]To derive equation (15) formally, begin by re-writing necessary condition (7) to make explicit the shadow value ($\delta(t)$) on the $F(t) \geq 0$ non-negativity constraint:

$$U'(F(t)) - \lambda\theta(t) - \phi(t) + \delta(t) = 0 \text{ where } F(t) \geq 0, \delta(t) \geq 0, F(t)\delta(t) = 0.$$

This condition, combined with the endpoint condition (14) for $\theta(t)$ and necessary condition (13), implies

$$\theta(t) = \int_t^{\infty} [U'(F(\tau)) + \delta(\tau)]e^{-(r+\lambda)(\tau-t)}d\tau$$

and, by implication, (15). To verify this result, differentiate the second displayed equation with respect to time and substitute out $\dot{\theta}(t)$ using (13) to obtain the first displayed equation.

Thus, whenever production is unconstrained (but non-zero), marginal utility rises in percentage terms at the discount rate as in the standard Hotelling model with zero extraction costs.[40]

From the perspective of an individual extraction firm taking the future oil price path as given, it is intuitive that the capacity constraint will bind whenever future prices will forever rise strictly slower than the rate of interest r.[41] A novel feature of our model is that capacity-constrained production can be optimal during periods when the oil price increases strictly faster than r, provided that the magnitude and duration of this increase are limited. To see this result, suppose that the oil price will rise strictly faster than r and then level off. In this case, the firm might be thought to have an incentive to shut in today and subsequently produce in a pulse all of the deferred production precisely when the future oil price is greatest in present value. However, the capacity constraint does not allow this arbitrage: any production that is deferred today cannot be completely recovered at the future instant it is most valuable.[42] Instead, the deferred production must be recovered over the full remaining life of the well, including both the time period when the oil price is higher than the current price in present value and the period when the oil price is lower than the current

[40]For an example in which unconstrained production is optimal, suppose that $K_0 > 0$, that there are no new wells remaining to be drilled, and that oil demand has a constant elasticity of $\eta > 0$ that is sufficiently small that if production declines exponentially at rate λ, marginal utility rises faster than r (i.e., $U'(F) = aF^{-\eta}$, where $\eta\lambda > r$). If production is always unconstrained, $\dot{U}'/U' = r$, requiring $F(t) = F(0)e^{-rt/\eta}$. This production program is optimal if all reserves are extracted in the limit and $F(t) \leq K(t)$ for all t. Complete extraction in the limit requires that $K_0/\lambda = F_0\eta/r$, which implies that $F_0 < K_0$. This argument applies at any given starting time t_0, so this production program is optimal.

[41]This statement can be proven by contradiction. Suppose that production is not constrained at some time t such that for all $\tau \geq t$, the oil price is rising strictly more slowly than r. Thus, $\phi(t) = 0$, implying (via condition (13)) that $\dot{\theta}(t)/\theta(t) = r$. Condition (7) then implies that, at least for some interval of time immediately following t, $F(\tau) = 0$ (since the oil price is rising strictly slower than r while $\theta(t)$ is rising at r, it must be that the oil price is strictly less than θ immediately after t, so that the complementary slackness condition then requires that oil flow equal zero). But then, with the oil price rising strictly slower than r forever, we must always have $F(\tau) = 0\ \forall\tau > t$, along with $\dot{\theta}(\tau)/\theta(\tau) = r\ \forall\tau > t$. But this result violates the transversality condition (14), a contradiction.

[42]More formally, suppose that production is reduced below the constraint by an amount $\epsilon > 0$ for a time interval of length $\delta > 0$. Then, the total amount of oil production deferred equals $\epsilon\delta$, and the available production capacity after this time interval will be $\lambda\epsilon\delta$ greater than it otherwise would have been. This additional capacity is not infinite, so the entire deferred volume cannot be extracted immediately. The fastest way to extract the deferred production is to produce at the capacity constraint, in which case the rate of production declines exponentially at rate λ, and the deferred production is only completely recovered in the limit as $t \to \infty$.

price in present value. If the value lost during the latter period outweighs the value gained during the former period, we will have $U'(F(t)) > \lambda\theta(t)$, so that it is optimal to operate at the capacity constraint even though the oil price will (temporarily) rise faster than r. The outcome of this tradeoff depends on the path of future prices relative to the current price and on the values of λ and r.

Our production data indicate that firms maintained production at the capacity constraint throughout our 1990–2007 sample. In section 4 below, we test whether our model can rationalize this behavior by empirically evaluating the $U'(F(t)) > \lambda\theta(t)$ inequality for 1990–2007, using crude oil spot prices to measure $U'(F(t))$ and using several measures of expected future oil prices to calculate $\theta(t)$. Many of these measures suggest that firms during 1998–1999 expected the price of oil to rise, for a time, faster than the rate of interest. The tests in section 4 therefore focus their attention on this period, when it is not clear *a priori* whether firms had an incentive to produce at the constraint.

3.3 Implications of the necessary conditions for drilling: a modified Hotelling rule

Condition (9) characterizes drilling incentives. The $\theta(t)X$ term is the value of the X units of capacity created by drilling a new well, and the $d(a(t))$ term is the marginal cost of drilling a well at time t when the rate of drilling is $a(t)$. The term $\gamma(t)$ can be interpreted as the shadow value of the marginal undrilled well at t. Therefore, besides the cost of renting a rig, a well owner who chooses to drill incurs an implicit cost of relinquishing his asset, the undrilled well. Condition (11) implies that $\gamma(t) = \gamma_0 e^{rt}$, where $\gamma_0 \geq 0$ is a constant. Intuitively, for someone to hold an undrilled well instead of some other asset, its value must rise by at least the discount rate and can rise no faster, or else agents would seek to purchase undrilled wells. Thus, when drilling occurs ($a(t) > 0$), conditions (9) and (11) together imply that the

marginal return to drilling must rise at r:

$$\theta(t)X - d(a(t)) = \gamma_0 e^{rt}. \tag{16}$$

Condition (9) is analogous to the standard Hotelling rule, which states that, constrained to a fixed volume of oil, the planner should extract so that the net marginal value of extracting *barrels* rises at the rate of interest. Thus, every barrel extracted yields the same net payoff in present-value terms. Since this is a drilling problem, however, the Hotelling-like intuition applies to *wells*, not to barrels.

Equation (16) holds whenever drilling occurs. If the right-hand side is strictly larger than the left-hand side even when $a(t) = 0$, it is optimal to refrain from drilling. A mundane way in which this situation may occur is if $d(0)$ is large relative to $U'(0)$. However, even if $d(0)$ is small relative to $U'(0)$, drilling must cease in finite time whenever marginal utility is bounded (as in the case of a substitutable backstop technology). For in that case, additional capacity cannot be worth more than $U'(0)/r$ even if the choke price were collected continuously. Hence, $e^{-rt}[\theta(t)X - d(0)] \to 0 < \gamma_0$ as $t \to \infty$.

During an interval while drilling is occurring, production may be either at the constraint or below it. It may at first seem counterintuitive that a planner would ever incur the cost of building additional capacity and then not use it immediately. But such a plan can be optimal if more capacity in the future is desirable and if, because $d(a)$ is increasing, it costs more to build that capacity over a short future interval when it will be fully utilized than over a longer period before it is utilized. Such advance preparations would never be optimal, of course, if one could somehow drill a mass of wells instantaneously in a pulse without raising the cost per well.

Whenever drilling proceeds over an interval where production is below capacity, the marginal cost of drilling must rise at the rate of interest; otherwise, one could rearrange the drilling path over that same interval and end with the same expansion of capacity but at

a lower discounted cost. Whenever drilling proceeds over an interval where production is capacity-constrained, however, the marginal cost of drilling must rise at less than the rate of interest. Although this result does mean that discounted drilling costs could be reduced by postponing drilling until later, doing so would result in an offsetting loss of utility since one would have to forgo some production (and, in a market context, some sales revenue) if the drilling were deferred.[43]

For the remainder of this section we focus on the empirically relevant case in which production is constrained during drilling. In this case, the necessary conditions can be manipulated to yield:[44]

$$U'(F(t)) - \left[\frac{(r+\lambda)d(a(t))}{X} - \frac{d'(a(t))\dot{a}(t)}{X}\right] = \frac{\lambda\gamma_0}{X}e^{rt}. \tag{17}$$

Equation (17) holds whenever production is constrained and $a(t) > 0$, and it can be interpreted as the modified per-barrel Hotelling rule for our model. On the right-hand side of (17) we have the shadow value of wells ($\gamma_0 e^{rt}$) divided by the total amount of oil stored in a unit mass of untapped wells (X/λ), which we can interpret as the per-barrel shadow value of oil in untapped wells. On the left-hand side is the marginal utility of oil less a term in square brackets reflecting the cost of the additional production. The first term in square brackets is the amortized, per-barrel marginal cost of drilling a well at time t.[45] The last term in square brackets captures the opportunity cost of drilling now versus waiting, which arises due to the

[43]To prove each of these results, note that whenever drilling occurs, equation (16) must hold. If the production constraint is not binding, $\theta(t)$ grows at the rate of interest, and this equation implies that $d(a(t))$ must also grow at that rate. On the other hand, if the production constraint binds, $\theta(t)$ grows at a slower rate and, therefore, so must $d(a(t))$.

[44]Derivation: If $F(t)$ is strictly positive, conditions (7) and (13) allow us to eliminate $\phi(t)$ from the system and obtain expression (i): $\dot{\theta}(t) = -U'(F(t)) + (r+\lambda)\theta(t)$. Differentiating condition (9) with respect to time and solving for $\dot{\theta}(t)$ yields expression (ii): $\dot{\theta}(t) = \frac{r\gamma(t)+d'(a(t))\dot{a}(t)}{X}$. Finally, we substitute for $\theta(t)$ and then $\dot{\theta}(t)$ in condition (16) using expressions (i) and then (ii), and assuming $a(t) > 0$, to yield equation (17).

[45]To clarify, consider the special case of constant marginal drilling costs: $d(a(t)) = \bar{d}$. If the planner drills one well (or rather, she marginally increases the rate of drilling) she obtains a marginal increase in oil flow h instants later of $Xe^{-\lambda h}$, assuming oil flow is set to the maximum. If each barrel of flow has imputed cost of c at that time, then at the time the well is drilled, the flow at h would have an imputed cost $cXe^{-(r+\lambda)h}$. Since such flows continue indefinitely, we want to find c such that $\bar{d} = \int_{h=0}^{\infty} cXe^{-(r+\lambda)h}dh$. Integrating, we obtain $\bar{d} = cX/(r+\lambda)$. Solving for c, we conclude: $c = \bar{d}(r+\lambda)/X$.

convexity in the drilling cost function. When $\dot{a}(t) < 0$, drilling activity and marginal drilling costs are falling over time, so that drilling immediately incurs an additional opportunity cost relative to delaying. One implication of this term is that, in contrast to standard Hotelling models in which production must decline over time (with the marginal extraction cost also declining if extraction costs are convex), oil flow in our model can *increase* over intervals during which the marginal cost of drilling is falling (i.e., $\dot{F}(t) > 0$ and $\dot{a}(t) < 0$ holding simultaneously is possible).

If drilling costs are affine rather than strictly convex, so that the last term in square brackets drops out of equation (17), then this equation becomes the Hotelling rule for a standard barrel-by-barrel extraction model with a constant marginal extraction cost of $(r + \lambda)d/X$. In fact, appendix D shows that if $d'(a) = 0$, and if $a(t) > 0$ and $F(t) = K(t)$ throughout the entire optimal path, then our model and the standard Hotelling model yield identical production and marginal utility paths.

A standard Hotelling model can therefore yield the same predictions as our reformulated model, but the conditions necessary for this equivalence are not realistic. First, equivalence requires the absence of unanticipated shocks along the equilibrium path. Oil production will immediately respond to such shocks in a standard Hotelling model, but the binding capacity constraint precludes this response in our model (as we discuss in section 6.2). Second, the assumption that $d'(a) = 0$ is clearly at odds with the data on rig rental prices shown in figure 2. In addition, in order to have that $a(t) > 0$ throughout the entire optimal path, it must be that $U'(F)$ is unbounded—an assumption which seems implausible given the likely emergence in the long run of alternative fuels and technologies.[46]

[46] Appendix D discusses two additional necessary conditions for this special case—K_0 must be sufficiently small and λ must be sufficiently large—that are more likely to be satisfied empirically.

4 Was producing at the constraint optimal in Texas?

Our data from Texas indicate that production from drilled wells declines steadily over time without substantially responding to the large swings in oil prices observed over 1990 to 2007. Can our model rationalize this behavior by showing that firms had an incentive to maintain production at its capacity constraint throughout this period?

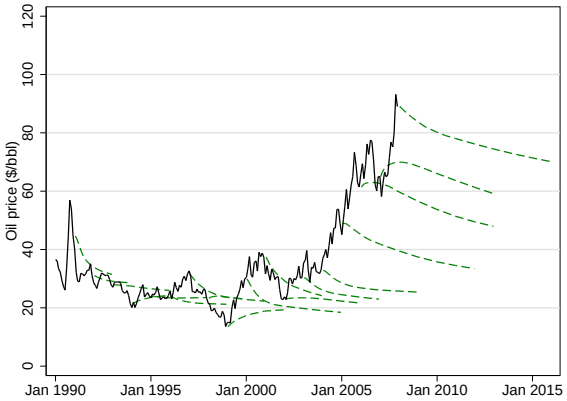
Our model yields a necessary condition for constrained production that in principle can be taken to data: $U'(F(t)) > \lambda\theta(t)$. On the left, the value generated by producing a barrel of oil today ($U'(F(t))$) is given by the spot price of oil. On the right, the value of deferring this production until some future date ($\lambda\theta(t)$) depends on the production decline rate while constrained (λ) and the shadow value of capacity ($\theta(t)$), which is determined by future oil prices and the discount rate (r) per equation (15). We set λ equal to 0.1, which is consistent with the observed production declines in Texas and stylized facts reported in Thompson (2001) and Mason and van't Veld (2013). We set $r = 0.1$, which consistent with a survey of oil producers during our sample period (see Society of Petroleum Evaluation Engineers 1995 and the discussion in Kellogg 2014). Measuring $\theta(t)$, however, is complicated by the fact that our model is deterministic, whereas real-world oil prices are stochastic. To build intuition for our approach to estimating $\theta(t)$, we begin in section 4.1 by using expected future prices directly in place of our model's deterministic price path. Then, in section 4.2, we re-calculate $\theta(t)$ while accounting for the option value that is generated by the volatility of real-world oil prices.

4.1 Estimation of the value of capacity using only expected future oil prices

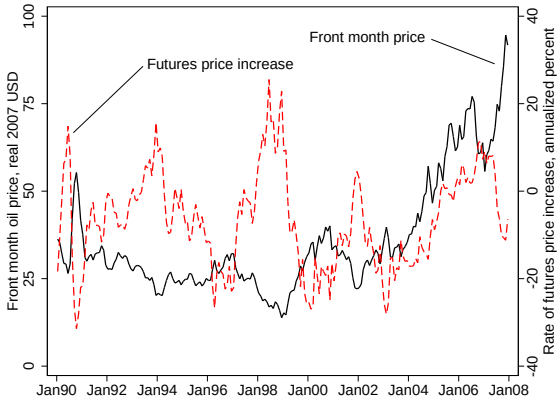
How should we measure oil producers' future price expectations? There is no clear correct answer, so in practice we consider several measures. An obvious starting point is to use NYMEX futures prices. As discussed in Alquist and Kilian (2010), futures prices are com-

Figure 3: Crude oil spot and futures prices

(a) Spot price and futures curves



(b) Spot price and rate of futures price increase



Note: In panel (a) the solid black line shows crude oil front month ("spot") prices, and the dashed lines show futures curves as of December in each year. Panel (b) shows the spot price and the percentage change in futures prices between the 12-month futures price and the spot price, for each month of the sample. All prices are real $2007, and the futures curves (panel a) and percentage changes in futures prices (panel b) are net of inflation. See text for details.

monly used by central banks and the International Monetary Fund to proxy for the market's price expectations. Moreover, a majority of industry participants claim to use futures prices in making their own price projections (Society of Petroleum Evaluation Engineers 1995). Finally, as we show in appendix A, we find that firms' on-lease above-ground oil stockpiles increase when futures prices exceed the spot prices by more than the rate of interest, as one would expect if firms' price expectations aligned with the futures market.

Figure 3(a) shows the time series of crude oil spot prices (solid line) as well futures curves as of December in each year (dashed lines).[47] For example, the left-most dashed line shows prices in December 1990 for futures contracts with delivery dates from January 1991 through

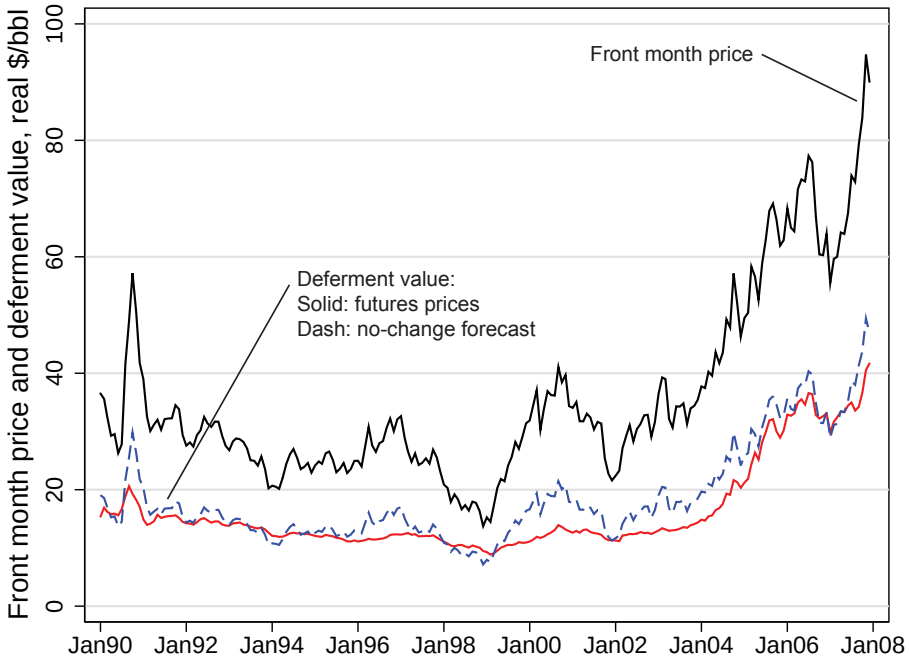
[47]We have converted all of the price data so that the slopes of the futures curves in figure 3 and the expected rates of price change used in our analysis reflect real rather than nominal changes. Our conversion accounts both for the trade date's CPI and for expected annualized inflation of 2.67% between the trade date and delivery date (the average annual inflation rate from January 1990 to December 2007 is 2.67%, and inflation varies little over the sample). For example, we convert the nominal prices for futures contracts traded in December 1990 to real price expectations by multiplying by the December 2007 CPI, dividing by the December 1990 CPI, and then dividing each contract price by $1.0267^{t/12}$, where t is the number of months between the trade date and the delivery date.

December 1992. The figure shows that the futures market for crude oil is often backwardated (meaning that the futures price is lower than the spot price) and was strongly backwardated during the mid-2000s when the spot price was rapidly increasing. Still, there were several periods of contango (meaning that the futures price is higher than the spot price) during the sample, particularly during 1998–1999 when the oil price was quite low. In fact, the 12-month ahead futures price during this period sometimes exceeded the spot price by more than 20% in real terms. This feature of the data is perhaps more clearly shown in panel (b) of figure 3, which shows the percentage difference (adjusted for inflation) between the 12-month futures price and the spot price for each month of the sample.

Taking futures prices as expected future prices, standard Hotelling models would predict that firms had a strong incentive to shut in production during 1998–1999. To test this incentive in our model, we calculate $\theta(t)$ for each month of the 1990–2007 sample, using futures prices and a discrete-time analog of (15). Our calculations, which we discuss in detail in appendix B.1, use a backward recursion procedure to account for the possibility that firms may find producing below the constraint to be optimal at some future date. For comparison, we also calculate $\theta(t)$ using a no-change forecast for oil prices, which Alquist and Kilian (2010) show to out-perform futures prices in out-of-sample forecasting over 1991–2007 (though Chernenko, Schwarz and Wright 2004 find the opposite for 1989–2003). Clearly, under a no-change forecast prices are never expected to rise faster than r, so the $U'(F(t)) > \lambda\theta(t)$ condition is guaranteed to hold.

Figure 4 plots the spot price of oil from 1990–2007 along with our calculations of the marginal value of deferred production ($\lambda\theta(t)$), based on both futures prices and a no-change forecast. The futures-based deferment value is below that based on a no-change forecast for most of the sample, owing to the frequent backwardation of futures prices, but it does exceed the no-change deferment value during periods such as 1998–1999 when the futures market is in contango. Nonetheless, for either measure of future price expectation, the value of deferred production never exceeds the spot price of oil, even during 1998–1999. This

Figure 4: Crude oil spot price vs. value of deferred production



Note: This figure shows the crude oil front month ("spot") price and the value of one barrel of deferred production ($\lambda\theta(t)$) in each month, all in real 2007 dollars. The time series showing the value of deferred production uses price expectations based on unadjusted futures prices and on a no-change forecast. See text and appendix B.1 for details.

finding is consistent with our production data showing that firms maintained production at the constraint throughout this period.

While futures prices serve as a convenient measure of firms' price expectations, Pindyck (2001) demonstrates that oil futures prices will generally not equal expected future prices of oil, owing to a risk premium. Intuitively, an agent who stores crude oil and takes a short position in the futures market owns a riskless portfolio that must in equilibrium yield a return equal to the riskless rate of interest. In contrast, an agent who simply stores crude oil without the hedge is exposed to the future oil price, which is stochastic. The expected return to simply storing crude oil, which depends on the expected future oil price, will then depend, in a CAPM framework, on the correlation between the oil price and the market portfolio. This difference in expected returns between hedged versus unhedged storage allows Pindyck

(2001) to formally relate the expected future price of oil to oil futures prices via equations (18) and (19):

$$E_t[P_{t+T}] = F_{t,T} + (r_{cT} - \rho_T)P_t \tag{18}$$

$$r_c = \rho + \beta(r_m - \rho) \tag{19}$$

In equation (18), P_t denotes the spot price, $E_t[P_{t+T}]$ denotes the expected spot price T periods ahead, and $F_{t,T}$ denotes the futures price for delivery T periods ahead. r_{cT} and ρ_T are the rates of return for crude oil and a riskless asset, respectively, at horizon T, and in equation (19) r_c and ρ denote their annualized values. The term r_m is the annualized market rate of return, and β is the CAPM beta for crude oil. A strictly positive value of β implies that un-hedged holders of crude oil bear non-diversifiable risk, so that the expected future price of oil must be strictly greater than the futures price.[48] Thus, if β is positive, our analysis above that calculates $\theta(t)$ using unadjusted futures prices will underestimate the value of deferring production.[49]

In appendix B.1, we show that a β of approximately 1.0 is required before producers have an incentive to defer production in even one month of the sample. We also show that, using NYMEX front-month oil prices and data on S&P 500 returns, we obtain a point estimate of

[48] A positive β is consistent with the idea that global demand and macroeconomic shocks outweigh oil supply shocks in determining crude oil prices, while the reverse is true for a negative β.

[49] The CAPM beta for crude oil is distinct from the beta for an oil producing firm—the former is relevant for assessing the difference between the expected future price of crude and the crude futures price, while the latter is relevant for assessing the firm's cost of capital. While we estimate a beta for crude oil that is nearly zero, other researchers (Kaplan and Peterson 1998; Sadorsky 2001; Damodaran 2015a) have estimated betas between 0.5 and 1.27 for firms in the oil and gas sector (in turn justifying discount rates for oil and gas firms around 10%, as used here and reported in Society of Petroleum Evaluation Engineers 1995). This difference in betas makes sense given that both positive oil demand shocks and positive oil supply shocks (e.g., technology improvements) will be associated with economy-wide growth and higher returns for oil and gas firms, leading to a strong correlation between oil and gas sector returns and broader market returns. However, these two types of shocks will have opposing effects on the price of oil, so that crude oil returns will not be strongly correlated with broader market returns. Our estimate of β for crude oil is also consistent with the only other estimate we were able to find in the literature: Scherer (2011) finds $\beta = 0.09$ using data from January 1997 to September 2008.

β equal to 0.05. Given this estimate, risk-adjusted futures prices will not appreciably differ from unadjusted futures prices, implying that our estimate of $\lambda\theta(t)$ will not appreciably differ from what is shown in figure 4, so that $U'(F(t)) > \lambda\theta(t)$ still holds. The standard error of our estimated β is 0.24, so that a 95% confidence interval strongly rejects $\beta = 1.0$. Thus, even after accounting for plausible values of the risk premium in futures prices, our model rationalizes why firms generally did not adjust oil production from existing wells in response to in-sample expectations that the oil price would temporarily rise faster than r.[50] Ideally, we would also like to be able to test whether producers shut in or at least reduced production at times when price expectations were such that $U'(F(t)) < \lambda\theta(t)$; however, such a circumstance was not realized during our sample.[51]

4.2 Estimation of the value of capacity allowing for stochastic oil prices

The main limitation of the above approach, which uses expected future prices directly to calculate $\theta(t)$ in our deterministic model, is that it ignores the fact that uncertainty about future oil prices generates option value for holding production capacity. That is, the option to use or not use capacity in the future—the exercise of which will depend on stochastic future price paths—enhances the value of holding capacity today.[52] Thus, our estimates of $\theta(t)$ from section 4.1 above are understated to the extent that this option value is important.

We account for option value in our calculation of $\theta(t)$ by recasting our model in a dynamic-stochastic optimization framework, as discussed in detail in appendix B.2. We begin by

[50] Alternatively, if one believed that producers had perfect foresight, they would have anticipated the dramatic increase in oil prices that occurred from 1999–2000 and would have had a strong incentive to shut in. However, a perfect foresight model is implausible in the context of historically volatile oil prices, and Alquist and Kilian (2010) note that forecasts based on industry surveys do not add predictive value over a no-change forecast.

[51] As we discuss in section 6.2 and in Anderson, Kellogg and Salant (2014), in equilibrium realizing $U'(F(t)) < \lambda\theta(t)$ requires a very large negative demand shock.

[52] This option value does not arise in standard Hotelling models in which the entire resource can be extracted in a pulse. In these models, the value of reserves is a function of expected future prices alone: it is the maximum of the discounted expected future prices.

econometrically estimating a first-order Markov process to characterize how both the crude oil spot price and the expected future price path evolve over time, using data on the evolution of both spot and futures prices. Our Markov process allows future price expectations to be either backwardated or in contango (and to stochastically switch between these states), possibly with an expected rate of price increase that exceeds r. We then computationally solve an infinite-horizon dynamic model of optimal production from any initial state (spot price and expected future price path), given our estimated Markov process. This model yields, for each month of our sample, an estimate of the marginal value of capacity ($\theta(t)$) that accounts for the option value associated with future price uncertainty. Parallel to our deterministic analysis from section 4.1, we first consider a stochastic price process in which expected future prices are given directly by futures market prices. We then adjust these expectations for non-diversifiable risk per the CAPM framework.

Measuring expected future prices using futures prices, we find that deferment values obtained from our stochastic model are uniformly higher than those from our deterministic approach from section 4.1. However, the difference never exceeds \$1 per barrel, so that the $U'(F(t)) > \lambda\theta(t)$ condition still holds. Intuitively, because episodes of severe contango are rare in the data, it is unlikely that oil producers will in the future want to shut in production, and the option value associated with holding capacity today is therefore small. When we measure expected prices using the CAPM adjustment, we find that deferment values rise, relative to those from the deterministic model, by a similar small amount. Thus, our conclusions based on the stochastic model are essentially the same as those from our relatively simple, deterministic analysis from section 4.1: producers in Texas never had an incentive to defer production, and a large CAPM beta exceeding 0.5 is still needed to overturn this conclusion.

4.3 Extending our model to account for fixed shut-in costs

If spot prices always exceeded the deferment value, then what explains the temporary shut-ins among marginal producers during the late 1990s? In appendix C, we extend our model to include a fixed production cost that must be incurred whenever a well's production is non-zero, and we show that fixed costs can rationalize shut-ins of marginal (low-volume) wells during 1998–1999 when the spot oil price was low but expected to rise in the future. We also show that the magnitude of fixed costs that rationalizes these shut-ins does not have a qualitatively important impact on the incentive to drill new wells, since the impact of fixed costs on the discounted revenue from a new well, θX, is small.[53]

5 Equilibrium dynamics with exogenous oil prices

Given our results on constrained production from section 4, the remainder of the paper focuses on drilling incentives and dynamics in cases where wells' oil production is always capacity-constrained. Throughout, we interpret our results in terms of a competitive equilibrium to relate our findings to data on oil prices and rig rental prices. We begin in this section by considering an individual oil field that is small relative to the global market, so that the path of oil prices is exogenous but the local rental market for drilling rigs clears at each instant. This case is particularly relevant for interpreting drilling and extraction behavior in a small, local region such as Texas. We treat the time path of oil prices $P(t)$ as given exogenously, while the path of rig rental rates ($d(a(t))$ is endogenously determined by the arbitrage condition (16) and the initial number of untapped wells (R_0). We defer consideration of equilibrium dynamics when oil prices are determined endogenously to section 6.[54]

[53] The intuition comes from the fact that shut-ins occur late in a well's life when it is producing only a few bbl/d, whereas new wells initially produce 100 bbl/d on average in our sample.

[54] Section 6 is "general" relative to the partial equilibrium analysis in section 5. However, like virtually all of the Hotelling literature, we do not model income effects generated by the rents earned by well owners (and in our case, by the owners of drilling rigs).

We focus on the case in which the oil price remains constant at $\bar{P}$ forever. In this case, per our discussion in section 3.2 and the proof in footnote 41, production is always capacity-constrained, and condition (15) holds with equality, implying that $\theta(t)X = \frac{\bar{P}X}{r+\lambda}$. Moreover, drilling activity must cease in finite time.

Substituting for $\theta(t)$ in equation (16), the arbitrage condition is given by:

$$d(a(t)) = \frac{\bar{P}X}{r+\lambda} - \gamma_0 e^{rt}. \tag{20}$$

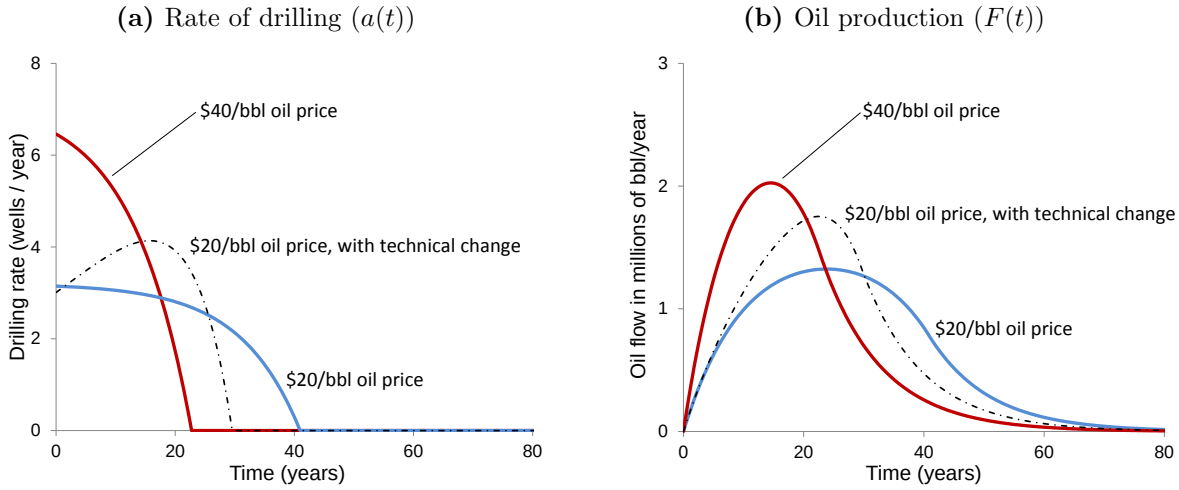
Equation (20) indicates that if drilling for oil is profitable ($\gamma_0 > 0$), then the equilibrium rig rental rate (the planner's marginal drilling cost) must decline over time so that drilling remains equally attractive as long as $a(t) > 0$. Because $d'(a) > 0$, the rate of drilling $a(t)$ must therefore also decline over time. Drilling must cease ($a(t) = 0$) when the initial supply R_0 of untapped wells is exhausted, and not beforehand.[55] This exhaustion condition uniquely determines the value of γ_0, the present value of an undrilled well,[56] so that the value of the undrilled stock of wells at $t = 0$ is given by $\gamma_0 R_0$. Equation (20) implies that the time of exhaustion T must satisfy $d(a(T)) = \bar{P}X/(r+\lambda) - \gamma_0 e^{rT}$, with $a(T) = 0$.

The solid lines in figure 5 illustrate the equilibrium time paths of drilling and production for a specific example starting with an initial capacity of $K_0 = 0$, an initial stock of undrilled wells $R_0 = 100$, and initial reserves per well (X/λ, more precisely reserves per unit mass of wells) of 0.5 million bbl. We use an affine drilling marginal cost curve given by $d(a) = 1 + 5a$ (with a in units of wells drilled per year and d in \$million per well), and we set $r = \lambda = 0.1$. Drilling activity ($a(t)$) follows the dynamics prescribed by equation (20), starting at a high level and declining until drilling ceases at T defined above, when all wells have been drilled. Oil production $F(t) = K(t)$ follows the dynamics prescribed by equation (12): $F(t)$ begins at zero, initially increases as the stock of drilled wells increases, but then eventually declines

[55] The intuition for this statement is obvious, but we supply a formal proof for the general equilibrium case in Anderson et al. (2014).

[56] To see this, observe that a very high value of γ_0 will result in the drilling of fewer than R_0 wells by the time T at which $a(T) = 0$. In contrast, a low value of γ_0 will cause the undrilled well stock R_0 to be exhausted at a time T for which $a(T) > 0$.

Figure 5: Optimal drilling in a local region



Note: This figure illustrates equilibrium time paths of drilling rates (panel a) and oil production (panel b), with exogenously given constant oil prices of \$20/bbl or \$40/bbl. The figure assumes an initial extraction capacity of $K_0 = 0$, an initial well stock of $R_0 = 100$, initial reserves per unit mass of wells of 0.5 million bbl, a decline rate of $\lambda = 0.1$, and a discount rate of $r = 0.1$. The solid lines assume a stationary marginal drilling cost curve given by $d(a) = 1 + 5a$, with a in units of wells drilled per year and d in \$million/well. The dashed line assumes that drilling costs are multiplied by $e^{-0.03t}$ and that this technological improvement is anticipated by firms.

asymptotically toward zero as the remaining volume of oil reserves (and reservoir pressure) declines. The resulting hump-shaped production profile is a well-known feature of production in oil fields across the world (Hamilton 2013). We show in section 6 below that a similar "peak oil" result also emerges in the case of endogenous oil prices.

Figure 5 also depicts how the equilibrium dynamics vary with the exogenous oil price $\bar{P}$. Because the total stock of wells is fixed, the oil price can affect the timing of drilling and production but not total cumulative drilling and production. For a relatively high oil price, the initial rate of drilling (and the rig rental rate) must be relatively high, and γ_0 must therefore also be relatively high to satisfy the exhaustion constraint.[57] Thus, a higher oil

[57] To see this, suppose γ_0 were unchanged (or reduced) from the low oil price case. In this case, equation (20) would require a uniformly higher rate of drilling over a longer period. This path would, therefore, call for more wells to be drilled than the available stock and cannot be an equilibrium. Increasing the value of γ_0 above its level in the low price case shortens the time interval over which wells are drilled so that the total number of wells drilled equals R_0.

price causes the stock of wells to be drilled relatively quickly, shifting oil production earlier in time (if the marginal drilling cost $d(a)$ is affine, as assumed in figure 5, it can be shown that peak production occurs sooner with a relatively high oil price). Moreover, because γ_0 is greater, a higher oil price increases the aggregate wealth derived from drilling the wells.

Note that this analysis comparing drilling, production, and rig rental paths under low and high price scenarios can be reinterpreted as an analysis of the responses of these variables to an unanticipated price shock. Suppose that the price of oil is anticipated to remain low forever but then, at t^*, suddenly and unexpectedly increases to a higher level, where it is then anticipated to persist. Drilling and rig rental rates immediately jump up at t^*, and production increases at a faster rate (or decreases at a slower rate) as the dynamics switch over to the drilling path corresponding to a higher oil price. Thus, the model replicates the covariance of oil prices, drilling, and rig rental rates that we observe in our Texas data.

Unanticipated shocks to the initial flow (X) or to the marginal cost of drilling additional wells can be analyzed in the same way. Indeed, since $\bar{P}$ and X enter multiplicatively in equation (20), doubling X will have the same effect on the paths of drilling and rig rental rates as doubling $\bar{P}$. Moreover, the wealth of well owners would jump up by the same amount in the two cases. Another shock that would induce this same drilling path would be an unanticipated halving of the marginal drilling cost associated with any drilling rate a due, say, to an unanticipated, cost-reducing technical change (well owners, however, would only be half as well-off after a shock that halves drilling costs as they would be after a shock that doubles the oil price).

Our analysis also clarifies the consequences of an unanticipated increase in reserves. If the supply of undrilled wells suddenly increases at some time t^*, then γ_{t^*} must immediately jump down, for otherwise drilling would cease before all the wells are drilled. Thus, the rate of drilling (and the rig rental rate) must jump up at t^*, and the interval over which $a(t)$ is non-zero will be longer than before.[58] This effect on the drilling path can be visualized in

[58]Similarly, a sequence of unanticipated increases in reserves would induce a sequence of upward jumps in drilling, with drilling declining monotonically after each upward jump because of the assumed stationarity

figure 5(a) as a rightward translation of the time path of drilling so that the drilling rate associated with the higher reserves is centered over t^*, the date when the unanticipated reserve increase occurred.

5.1 Impacts of steady, anticipated technological progress

Since 2009, there have been substantial increases in the productivity of shale oil extraction via hydraulic fracturing. The shale boom has been characterized by the rapid emergence of new technology that has been adopted by a large number of firms across multiple regions of the United States.[59] Figure 6 shows how shale oil has dramatically increased both U.S. drilling rig activity and oil production since 2009.[60] From 2010 through mid-2014, both oil drilling and production steadily increased even though oil prices were roughly constant.

Drilling can never steadily increase in our fixed-price model, since we assumed the marginal drilling cost function was stationary. To account for the persistent technological progress that resulted from the development of hydraulic fracturing, we augment our model to allow the marginal cost of drilling to shift down at an exogenous rate, and we assume that well owners anticipate these shifts.[61] Suppose that the marginal cost of drilling at any given drilling rate decays exponentially at rate μ over time. The modified arbitrage

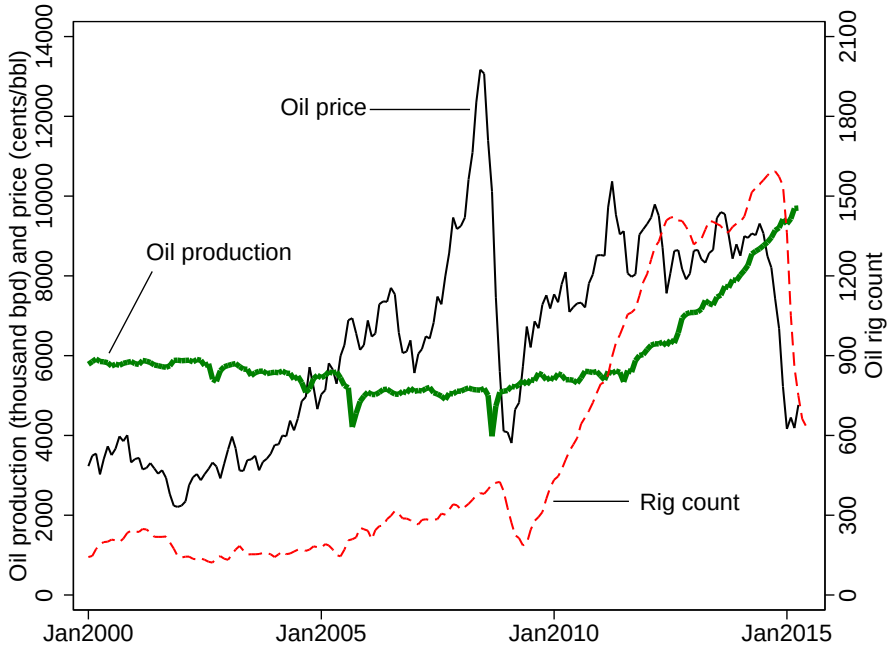
of the marginal cost function.

[59]For this reason, we treat the shale boom as a series of marginal cost reductions resulting from exogenous, technological change rather than from learning-by-doing, although both mechanisms have been at work to different extents. A detailed discussion of productivity improvements in the Bakken Shale in North Dakota is given in Covert (2015). In results available upon request, it can be shown that a learning-by-doing model is capable of generating a steady increase in the rates of both drilling and production, given certain parameter values and functional forms.

[60]Figure 6 shows oil drilling rig counts (the number of rigs actively drilling at any point in time) rather than the number of oil wells drilled because there is no aggregate database of U.S. oil well drilling.

[61]Steady *unanticipated* technology improvements can also generate an increasing rate of drilling (the drilling rate will actually increase more quickly in such a model), but we believe that anticipated technological progress is a more natural assumption for the shale boom, at least once the boom began.

Figure 6: U.S. oil rig count, oil production, and oil price



Note: U.S. monthly oil production data cover all 50 states and come from EIA. Front month ("spot") oil prices are real $2007 and also come from EIA. Monthly counts of rigs drilling for oil come from Baker-Hughes.

condition is then:[62]

$$\frac{\bar{P}X}{r+\lambda} - d(a(t))e^{-\mu t} = \gamma_0 e^{rt}. \tag{21}$$

If we take the derivative of equation (21) with respect to time and solve for $\dot{a}(t)$, we obtain

$$\dot{a}(t) = [d'(a(t))]^{-1}\mu e^{\mu t}\left[\frac{\bar{P}X}{r+\lambda} - \frac{\mu+r}{\mu}\gamma_0 e^{rt}\right], \tag{22}$$

the sign of which is dictated by the term in brackets. Thus, the drilling rate increases over time if and only if the revenue earned by a newly drilled well ($\bar{P}X/(r+\lambda)$) exceeds $(\mu+r)/\mu$ times the value of an undrilled well ($\gamma_0 e^{rt}$).

Note that for μ sufficiently large (i.e., rapid technological progress), the bracketed portion

[62] Our results qualitatively generalize to decay functions other than the exponential. Moreover, the steady reduction in drilling costs is isomorphic to a steady increase in reserves per well. Multiplying X by $(1+(r+\lambda)d(a(t))(1-e^{-\mu t})/(\bar{P}X))$ yields the same drilling and production paths as does equation (21) (though the value of γ_0 is different).

evaluated at $t = 0$ must be positive, such that drilling and production will initially expand simultaneously.[63] For t sufficiently large, however, the bracketed portion turns negative and drilling declines monotonically thereafter: $\dot{a}(t) < 0$. Intuitively, sufficiently rapid downward shifts in the drilling supply curve initially lead drilling activity to increase, even as the value of an undrilled well (i.e., the scarcity rent) grows at the rate of interest. However, drilling costs are bounded from below by zero. Thus, the relentlessly growing scarcity rent must eventually dominate, driving a decline in drilling. Meanwhile, for low values of μ (i.e., slow technological progress), the rising scarcity rent always dominates, so the drilling rate never increases.

To illustrate these results, the dashed lines in figure 5 present drilling and production rates obtained from a simulation in which we apply a 3% annual rate of technical progress ($\mu = 0.03$) and leave all other parameters unchanged. Both drilling and production increase steadily over the early years of the simulation before resource scarcity ultimately drives a decline in drilling. Thus, our model demonstrates that sustained technological progress can indeed rationalize the simultaneous increases in drilling and production that have occurred during the recent U.S. shale boom.

6 Equilibrium dynamics with endogenous oil prices

We now close the model by endogenizing the path of oil prices and requiring that at every instant demand equals supply. To ensure that the time paths for production, drilling, and oil prices are well-behaved, we assume that inverse oil demand ($P(F)$) is strictly decreasing, that drilling supply ($d(a)$) is strictly increasing, that both functions are continuously differentiable, and that $d(0)$ is sufficiently small relative to $P(0)$ that drilling occurs in equilibrium (i.e., $a(t) > 0$ for at least some t).[64]

[63] The first term in brackets must exceed the second term for μ sufficiently large and $t = 0$ because $(\mu + r)/\mu \to 1$ as $\mu \to \infty$ and γ_0 is bounded from above by $\frac{\bar{P}X}{r+\lambda}$.

[64] We prove in Anderson et al. (2014) that these conditions are sufficient for the continuity over time of production, drilling, and oil prices, and sufficient for other desirable properties such as the exhaustion of all

In section 5, we showed that starting from an initial capacity of $K_0 = 0$, the rate of drilling under an exogenous, fixed oil price of $\bar{P}$ is initially high but decreases monotonically over time. Intuitively, a similar result will typically hold when oil prices are endogenously determined. Starting from $K_0 = 0$, we initially have $F_0 = 0$ so that the oil price is at its upper bound, generating a strong incentive to drill at a rapid rate at $t = 0$. From there, the oil price must initially decline as capacity, and therefore production, is added. This price decline will cause $\theta(t)$ to decrease subsequent to $t = 0$. Thus, in order to have the current value of an undrilled well increase at r per equation (16), the marginal drilling cost and therefore the rate of drilling must initially be decreasing.

In Anderson et al. (2014), we show that with a weak condition on the shape of the demand curve, the drilling rate starting from K_0 sufficiently small will be decreasing not just initially but throughout the entire equilibrium path, yielding a single-peaked production profile.[65] We also show that under this same condition, production will be capacity-constrained along the entire equilibrium path.

On the other hand, if the initial capacity endowment K_0 is sufficiently large relative to oil demand, the initial rate of drilling with endogenous prices will be low, owing to the initially low oil price.[66] As capacity is exhausted, the oil price will rise, which can cause the drilling rate to initially increase over time (though it must eventually fall as the stock of undrilled wells nears exhaustion). For very large capacity endowments, production may initially be unconstrained.

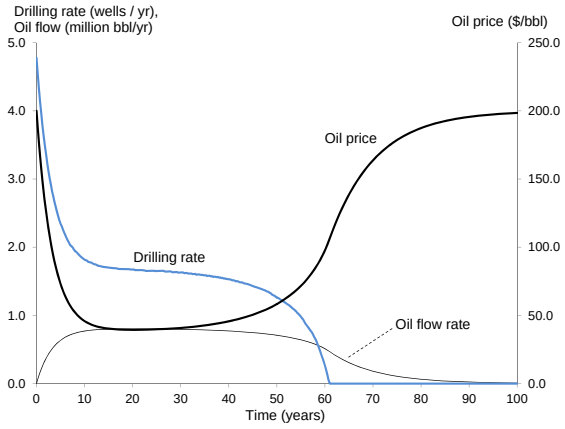
capacity and the drilling of all wells in the limit.

[65]The sufficient condition is that the inverse demand elasticity be weakly increasing in F. In this case, the incentive to defer production decreases over time as F decreases. Thus, once oil production and the rate of drilling are both decreasing over time, there is no incentive to increase the drilling rate. Moreover, once production is capacity-constrained (as it must initially be if K_0 is small), it will then always be constrained.

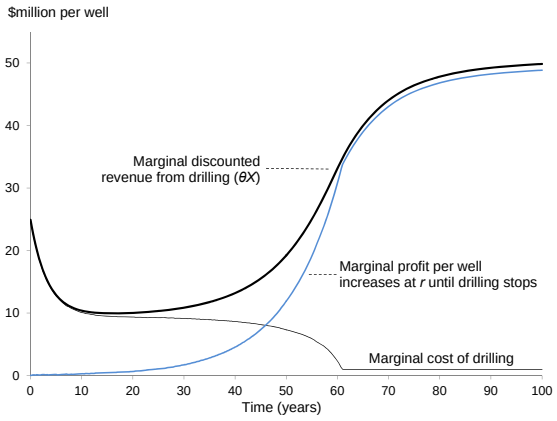
[66]Historically, we must of course have $K_0 = 0$. However, a sufficiently large negative demand shock along the original path would effectively put the system on a new equilibrium path corresponding to a large capacity endowment at time zero.

Figure 7: Equilibrium paths with linear demand and marginal drilling costs

(a) Drilling, production, and oil price



(b) Drilling incentives



Note: This figure illustrates the equilibrium time paths of drilling, oil production, and prices for a model with a linear inverse demand curve given by $P(F) = 200 - 200F$ (F in million bbl/d) and marginal drilling costs given by $d(a) = 1 + 5a$ (with d in million \$/well and a in wells/year). The initial well stock is $R_0 = 100$, with 0.5 million barrels of reserves per unit mass of wells, and $r = \lambda = 0.1$. See text and appendix F for details.

6.1 A specific example with linear demand and marginal drilling costs

Figure 7 depicts equilibrium dynamics for a specific example with parameters equal to those used in our exogenous price model in figure 5, except that firms now face a downward sloping inverse demand curve given by $P(F) = 200 - 200F$ (F in million bbl/d). This model satisfies the sufficient conditions in Anderson et al. (2014), so that production is always capacity-constrained starting from $K_0 = 0$. Because the model does not permit an analytic solution, we solve for equilibrium dynamics computationally using value function iteration; see appendix F for details.

Panel (a) illustrates that the rate of drilling is initially high but decreases rapidly. The rate of decrease slows to near-zero between years 10 and 40, but then accelerates so that all drilling is completed in year 61. Thus, the rate of oil production rapidly rises from zero at $t = 0$ to a plateau-like peak and then enters a decline as the drilling rate falls to zero. The

path of oil prices is therefore U-shaped.

Panel (b) depicts the time path of drilling incentives. The marginal discounted revenue from drilling $\theta(t)X$ initially falls as the oil price falls but then rises, asymptotically approaching \$50 million per well (equal to the asymptotic oil price of \$200/bbl multiplied by $X = 0.05$ million bbl/d and divided by $r + \lambda$). Marginal drilling costs $d(a(t))$ fall with the rate of drilling so that the marginal profit per well (equal to $\theta(t)X - d(a(t))$) rises at the interest rate until drilling ceases, adhering to necessary condition (16).

6.2 Impacts of unanticipated demand shocks

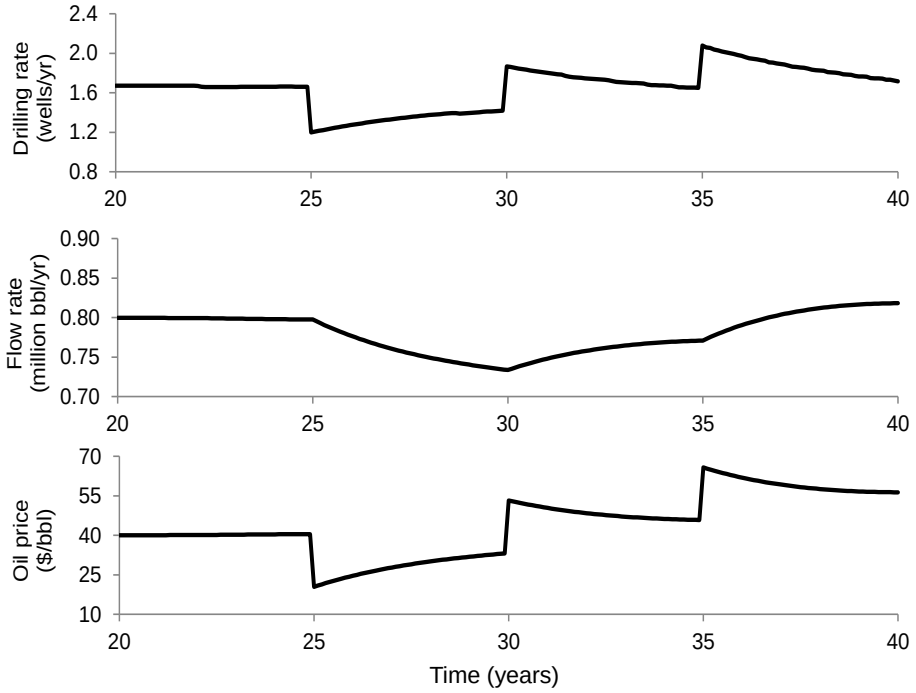
In this section, we explore how persistent, unanticipated shocks to oil demand affect drilling, production, and prices in equilibrium. Our motivation derives from the large historical changes in oil prices shown in figure 3, which prior work has attributed primarily to demand shocks (Kilian 2009; Kilian and Hicks 2013).

Consider a capacity-constrained equilibrium path and an unexpected, permanent, vertical shift in $P(F)$ of magnitude Z. This shift will cause the rate of drilling to increase immediately on impact,[67] implying that the rental rate for drilling rigs will also increase. Production, however, cannot immediately increase because it is capacity-constrained; thus, the oil price must increase on impact. If the shock is large enough, the jump up in drilling will be sufficient to cause production to gradually increase as new wells come on line, causing prices to gradually fall subsequent to the initial impact (a small shock may not cause production to rise if production was falling prior to the shock).

By the opposite mechanism, a negative demand shock will cause the oil price, rate of drilling, and rig rental rate all to decrease immediately on impact. Subsequently, production may fall (if the shock was large enough or production was falling already), leading the oil price to gradually rise after impact. For a sufficiently large shock, the price may rise faster

[67] Proof: suppose by contradiction that there is no change in the path of drilling. In this case, the entire path for $\theta(t)$ will increase by $Z/(r+\lambda)$, per equation (15). This shift implies that, under the original drilling path, $\theta(t)X - d(a(t))$ rises more slowly than the rate of interest, which is sub-optimal.

Figure 8: Impacts of unexpected demand shocks from the equilibrium model with linear demand, linear marginal drilling costs, and a finite well stock



Note: This figure uses the demand, drilling cost, and parameter specifications from figure 7. At time $t = 25$, demand is decreased from $P = 200 - 200F$ to $P = 180 - 200F$. At $t = 30$, demand is increased to $P = 200 - 200F$. Finally, at $t = 35$, demand is increased to $P = 220 - 200F$. See text for details.

than the interest rate, provided that oil demand is sufficiently inelastic.[68]

To illustrate these predicted responses, we expose the numerical model used to generate figure 7 to a series of three demand shocks: one negative shock followed by two positive shocks. The results of this simulation are shown in figure 8. The top panel shows that drilling "jumps" on impact for each shock. The middle panel illustrates that oil production does not jump following any shock but rather responds gradually to changes in the rate of drilling. The bottom panel shows that after the negative demand shock, the oil price jumps down on impact but then gradually rises, and then after the positive demand shocks the oil

[68]For an extremely large negative demand shock, production may actually fall below the capacity constraint upon impact, with the oil price and marginal cost of drilling both subsequently rising at the rate of interest. The fact that oil production from previously drilled wells does not respond to shocks during our 1990–2007 sample suggests that negative oil demand shocks during this period were not large enough to induce unconstrained production. The analysis underlying figure 4 confirms this result.

price jumps up but then gradually falls.

These simulation results are consistent with our empirical evidence from section 2 (and Kilian 2009 and Kilian and Hicks 2013), where demand shocks have immediate impacts on the oil price, drilling, and rig rental rate but not on production. Our results also help to shed light on the behavior of oil futures markets following demand shocks. After the simulated negative demand shock, our model indicates that the decrease in drilling activity should lead to a gradual reduction in production and therefore a gradual rise in the oil price. Thus, upon impact of the shock, market participants should anticipate this future price rise (in the absence of any subsequent shocks). Turning to the futures market data in figure 3, we see that following the negative demand shock from the Asian financial crisis in 1998–1999 (Kilian 2009), oil futures prices were in strong contango, suggesting that market participants indeed believed that prices would subsequently rise.[69] Conversely, following the positive global demand shocks in the mid-2000s, oil futures markets were backwardated. This backwardation suggests a belief that prices would fall in the absence of further shocks, and our model's simulation results suggest that an anticipated supply response could naturally lead market participants to hold such beliefs.

Of course, our comparisons between futures markets and simulated price paths are not precise: the real-world oil market is continuously affected by shocks, while our model is deterministic so that agents never expect any shocks. To yield predictions that could quantitatively be taken to the data, our model would need to be extended to allow for unanticipated demand shocks in each period, and we would need to solve the planner's problem (1) in a stochastic environment. In this case, each undrilled well would be characterized as a real option (see Kellogg 2014).

[69]Recall that we find in section 4 that the risk adjustment required to translate futures prices to price expectations is quite small in a CAPM framework, so that futures prices are a reasonable measure of firms' future price expectations.

7 Conclusion

Most Hotelling-style models take for granted that extractors are unconstrained in choosing the rate at which production flows to market. Our analysis of crude oil drilling and production in Texas shows this assumption to be inconsistent with the technology and cost structure of the oil industry. Oil is not extracted barrel-by-barrel. Instead, extractors drill wells, and the maximum flow from these wells is geologically constrained by underground reservoir pressure. Given this constraint, extractors always opt to produce at capacity under realistic conditions and will respond to oil prices only by varying their rate of drilling. Thus, crude oil extraction is best modeled as a dynamic drilling investment problem in which the flow of oil—though technically a control variable—is set equal to production capacity at all times and therefore behaves like a state variable in equilibrium.

We develop a new model of exhaustible resource extraction that accommodates these important features of the crude oil extraction industry. Our model—uniquely in the Hotelling literature—replicates several salient observations from oil markets: (1) production from pre-existing wells steadily declines over time and does not respond to oil price shocks, even when the oil price is anticipated to rise (temporarily) faster than the rate of interest; (2) drilling of new wells and drilling rig rental rates strongly co-vary with oil prices; (3) local oil-producing regions and fields exhibit production peaks; (4) steady technological progress in a local region can cause the rates of drilling and production to both steadily increase; and (5) following an unanticipated positive demand shock, the oil price will jump up on impact but can then gradually fall (with the reverse holding for an unanticipated negative demand shock). A main contribution of our paper is therefore to show that a Hotelling-style model *can* replicate qualitatively the oil price, extraction, and drilling dynamics that we observe in the real world, breathing new life into a theoretical literature that has, until now, largely failed to deliver empirically.

Our model could be extended in several logical ways. First, we currently model drilling costs with a fixed, upward-sloping supply curve. While the stock of drilling rigs and crews is

fixed in the short run, with rigs allocated to their highest-valued use, this stock can change over time as new rigs are built and crews are trained, as old rigs are scrapped and workers retire, or as rigs and crews are moved from one region to another. Market power by rig owners may also be a factor in rig investment and utilization decisions. These rig dynamics could be added to the model to effectively allow for a more elastic long-run drilling supply curve. Second, the model could incorporate uncertainty about future oil demand or drilling costs, so that firms rationally expect shocks to occur along the equilibrium path. Third, since different locations typically have their own geological features, it would be natural to consider variation in drilling costs, production decline rates, and resource stocks across regions—or across individual wells within the same region. Heterogeneity in drilling costs may be particularly important to consider given the increasing reliance on deeper, more remote, and unconventional energy resources. We leave these extensions to future work.

References

Adelman, M. A., "Mineral Depletion, with Special Reference to Petroleum," *The Review of Economics and Statistics*, 1990, *72* (1), 1–10.

Alquist, Ron and Lutz Kilian, "What Do We Learn from the Price of Crude Oil Futures?," *Journal of Applied Econometrics*, June/July 2010, *25* (4), 539–573.

Anderson, Soren T., Ryan Kellogg, and Stephen W. Salant, "Hotelling Under Pressure," 2014. NBER Working Paper 20280.

Barnett, H. J. and C. Morse, *Scarcity and Growth*, Baltimore, MD: Johns Hopkins University Press, 1963.

Berck, Peter and Michael Roberts, "Natural Resource Prices: Will They Ever Turn Up?," *Journal of Environmental Economics and Management*, 1996, *31*, 65–78.

Black, Geoffrey and Jeffrey T. LaFrance, "Is Hotelling's Rule Relevant to Domestic Oil Production?," *Journal of Environmental Economics and Management*, 1998, *36*, 149–169.

Boomhower, Judson, "Drilling like There's No Tomorrow: Bankruptcy, Insurance, and Environmental Risk," 2015. Unpublished manuscript, accessed from `https://dl.dropboxusercontent.com/u/14325430/Boomhower_07112015.pdf` on 29 September, 2015.

Cairns, Robert, "The Green Paradox of the Economics of Exhaustible Resources," *Energy Policy*, February 2014, *65*, 78–85.

Cairns, Robert D. and Graham A. Davis, "Adelman's Rule and the Petroleum Firm," *The Energy Journal*, 2001, *22* (3), 31–54.

Chernenko, Sergey V., Krista B. Schwarz, and Jonathan H. Wright, "The Information Content of Forward and Futures Prices: Market Expectations and the Price of Risk," 2004. Board of Governors of the Federal Reserve System, International Finance Discussion Paper 808.

Covert, Thomas R., "Experiential and Social Learning in Firms: The Case of Hydraulic Fracturing in the Bakken Shale," 2015. Unpublished manuscript, accessed from `http://home.uchicago.edu/~tcovert/webfiles/fracking.pdf` on 9 August, 2015.

Dahl, Carol and Mine Yucel, "Testing Alternative Hypotheses of Oil Producer Behavior," *The Energy Journal*, 1991, *12* (4), 117–138.

Damodaran, Aswath, "Betas by Sector (US)," 2015a. Accessed from `http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/Betas.html` on 6 June, 2015.

_ , "Equity Risk Premiums (ERP): Determinants, Estimation, and Implications—the 2015 Edition," 2015b. Unpublished manuscript, http://ssrn.com/abstract=2581517.

Darcy, Henry, *Les Fontaines Publiques de la Ville de Dijon*, Paris: Victor Dalmont, 1856.

Davis, Graham A. and Robert D. Cairns, "Simple Analytics of Valuing Producing Petroleum Reserves," *The Energy Journal*, 1998, *19* (4), 133–142.

Denver Post, "Economy unearths stripper well risks," 2008. Accessed from `http://www.denverpost.com/business/ci_10990024` on 24 July, 2015.

Devarajan, Shantayanan and Anthony C. Fisher, "Hotelling's 'Economics of Exhaustible Resources': Fifty Years Later," *Journal of Economic Literature*, 1981, *19* (1), 65–73.

Elliott, Graham, Thomas J. Rothenberg, and James H. Stock, "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, July 1996, *64* (4), 813–836.

Farrow, Scott, "Testing the Efficiency of Extraction from a Stock Resource," *Journal of Political Economy*, 1985, *93* (3), 452–487.

Financial Times, "US stripper well operators eye closures amid low oil price," 2014. Accessed from `http://www.ft.com/cms/s/0/f72c4c2c-8340-11e4-b017-00144feabdc0.html\#axzz3gq7w1w28` on 24 July, 2015.

Gao, Weiyu, Peter R. Hartley, and Robin C. Sickles, "Optimal Dynamic Production from a Large Oil Field in Saudi Arabia," *Empirical Economics*, 2009, *37* (1), 153–184.

Gaudet, Gérard, "Optimal Investment and Adjustment Costs in the Economic Theory of the Mine," *Canadian Journal of Economics*, February 1983, *16* (1), 39–51.

_ , "Natural Resource Economics Under the Rule of Hotelling," *Canadian Journal of Economics*, 2007, *40* (4), 1033–1059.

_ **and Stephen Salant**, "The Hotelling Model with Multiple Demands," in Robert Halvorsen and David Layton, eds., *Handbook on the Economics of Natural Resources*, Edward Elgar, 2014.

Ghandi, Abbas and C.-Y. Cynthia Lin, "An Analysis of the Economic Efficiency of Iraqs Oil Service Contracts," 2014. UC Davis Working Paper.

Griffin, James M, "OPEC Behavior: A Test of Alternative Hypotheses," *American Economic Review*, December 1985, *75* (5), 954–963.

Güntner, Jochen H. F., "How Do Oil Producers Respond to Oil Demand Shocks?," *Energy Economics*, forthcoming.

Halvorsen, Robert and Tim R. Smith, "On Measuring Natural Resource Scarcity," *Journal of Political Economy*, 1984, *92* (5), 954–964.

_ **and** _ , "A Test of the Theory of Exhaustible Resources," *Quarterly Journal of Economics*, 1991, *106* (1), 123–140.

Hamilton, James D., "Understanding Crude Oil Prices," *The Energy Journal*, 2009, *30* (2), 179–206.

_ , "Oil Prices, Exhaustible Resources, and Economic Growth," in Roger Fouquet, ed., *Handbook of Energy and Climate Change*, Cheltenham, United Kingdom: Edward Elgar Publishing, 2013.

Hogan, William W., "A Dynamic Putty–Semi-putty Model of Aggregate Energy Demand," *Energy Economics*, 1989, *11* (1), 53–69.

Holland, Stephen P., "Modeling Peak Oil," *The Energy Journal*, 2008, *29* (2), 61–79.

Hotelling, Harold, "The Economics of Exhaustible Resources," *Journal of Political Economy*, 1931, *39* (2), 137–175.

Hyne, Norman J., *Nontechnical Guide to Petroleum Geology, Exploration, Drilling, and Production*, 2nd ed., Tulsa, Oklahoma: PennWell Corporation, 2001.

Jones, Clifton T., "OPEC Behaviour Under Falling Prices: Implications for Cartel Stability," *The Energy Journal*, 1990, *11* (3), 117–130.

Kaplan, Paul D. and James D. Peterson, "Full-Information Industry Betas," *Financial Management*, 1998, *27*, 85–93.

Kellogg, Ryan, "Learning by Drilling: Interfirm Learning and Relationship Persistence in the Texas Oilpatch," *The Quarterly Journal of Economics*, 2011, *126* (4), 1961–2004.

_ , "The Effect of Uncertainty on Investment: Evidence from Texas Oil Drilling," *American Economic Review*, 2014, *104* (6), 1698–1734.

Kilian, Lutz, "Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market," *American Economic Review*, 2009, *99* (3), 1053–1069.

_ **and Bruce Hicks**, "Did Unexpectedly Strong Economic Growth Cause the Oil Price Shock of 2003–2008?," *Journal of Forecasting*, August 2013, *32*, 385–394.

Krautkraemer, Jeffrey A., "Nonrenewable Resource Scarcity," *Journal of Economic Literature*, 1998, *36* (4), 2065–2107.

Léonard, Daniel and Ngo van Long, *Optimal Control Theory and Static Optimization in Economics*, Cambridge: Cambridge University Press, November 1992.

Lewis, Eric, "Federal Environmental Protection and the Distorted Search for Oil and Gas," 2015. Unpublished manuscript, accessed from `http://www-personal.umich.edu/~erickyle/eric_lewis_JMP.pdf` on 29 September, 2015.

Livernois, John R. and Russell S. Uhler, "Extraction Costs and the Economics of Nonrenewable Resources," *Journal of Political Economy*, 1987, *95* (1), 195–203.

Mason, Charles F. and Klaas van't Veld, "Hotelling Meets Darcy: A New Model of Oil Extraction," June 2013. Paper presented at the 20th Annual Conference of the European Association of Environmental and Resource Economists (Toulouse, France).

Miller, Merton H. and Charles W. Upton, "A Test of the Hotelling Valuation Principle," *Journal of Political Economy*, 1985, *93* (1), 1–25.

Muehlenbachs, Lucija, "A Dynamic Model of Cleanup: Estimating Sunk Costs in Oil and Gas Production," *International Economic Review*, 2015, *56* (1), 155–185.

_ , Elisheba Spiller, and Christopher Timmins, "The Housing Market Impacts of Shale Gas Development," *American Economic Review (forthcoming)*, 2015.

Nystad, Arild N., "Rate Sensitivity and the Optimal Choice of Production Capacity of Petroleum Reservoirs," *Energy Economics*, 1987, *9* (1), 37–45.

Okullo, Samuel J., Frédéric Reynès, and Marjan W. Hofkes, "Modeling Peak Oil and the Geological Constraints on Oil Production," *Resource and Energy Economics*, 2015, *40*, 36–56.

Pindyck, Robert S., "The Optimal Exploration and Production of Nonrenewable Resources," *Journal of Political Economy*, October 1978, *86* (5), 841–861.

_ , "The Dynamics of Commodity Spot and Futures Markets: A Primer," *Energy Journal*, 2001, *22* (3), 1–29.

Ramcharran, Harri, "Oil Production Responses to Price Changes: An Empirical Application of the Competitive Model to OPEC and Non-OPEC Countries," *Energy Economics*, 2002, *24* (2), 97–106.

Rao, Nirupama S., "Taxation and the Extraction of Exhaustible Resources: Evidence from California Oil Production," April 2010. MIT-CEEPR Working Paper 10-006.

Reuters, "U.S. oil well shut-ins start as crude rout batters small producers," 2015. Accessed from `http://www.reuters.com/article/2015/01/27/us-oil-wells-idUSKBN0L00CW20150127` on 24 July 2015.

RigData, "Day Rate Report," 1990–2013. Accessed from `http://www.rigdata.com/news-and-analysis/day-rate-report.aspx` on 27 August, 2013.

Sadorsky, Perry, "Risk Factors in Stock Returns of Canadian Oil and Gas Companies," *Energy Economics*, 2001, *23*, 17–28.

Salant, Stephen W., "The Economics of Natural Resource Extraction: A Primer for Development Economists," *World Bank Research Observer*, February 1995, *10* (1), 93–111.

Scherer, Bernd, "Portfolio Choice for Oil-Based Sovereign Wealth Funds," *The Journal of Alternative Investments*, Winter 2011, *13* (3), 24–34.

Slade, Margaret E., "Trends in Natural Resource Commodity Prices: An Analysis in the Time Domain," *Journal of Environmental Economics and Management*, 1982, *9*, 122–137.

_ **and Henry Thille**, "Whither Hotelling: Tests of the Theory of Exhaustible Resources," *Annual Review of Resource Economics*, August 2009, *1* (1), 239–259.

Smith, James L., "Modeling the Impact of Taxes on Petroleum Exploration and Development," 2012. IMF Working Paper No. 12/278.

Smith, V. Kerry, "Natural Resource Scarcity: A Statistical Analysis," *Review of Economics and Statistics*, 1979, *61*, 423–427.

Society of Petroleum Evaluation Engineers, "Survey of Economic Parameters Used in Property Evaluations," 1995.

Stollery, Kenneth R., "Mineral Depletion with Cost as the Extraction Limit: A Model Applied to the Behavior of Prices in the Nickel Industry," *Journal of Environmental Economics and Management*, February 1983, *10*, 151–165.

Switzer, Sheldon and Stephen Salant, "Optimal Capacity Expansion by Foresighted Extractors and Producers," in G. Gaudet and P. Lasserre, eds., *Problèmes de Ressources Naturelles et Théorie Économique*, University of Laval Press, Laval, Québec, 1986.

Thompson, Andrew C., "The Hotelling Principle, Backwardation of Futures Prices and the Values of Developed Petroleum Reserves—the Production Constraint Hypothesis," *Resource and Energy Economics*, 2001, *23*, 133–156.

Venables, Anthony J., "Depletion and Development: Natural Resource Supply with Endogenous Field Opening," *Journal of the Association of Environmental and Resource Economists*, 2014, *1* (3), 313–336.

Wood Mackenzie, "Analysis of 2,222 oil fields suggests if Brent falls to $40 a barrel 1.6% of global oil supply could be cash negative," 2015. Accessed from `http://www.woodmac.com/public/views/12525771` on 24 July 2015.

A Additional empirical results

This empirical appendix has three parts. First, we provide results from regressions that complement figures 1 and 2 in the main text, and we show how re-entries co-vary over time with the oil price. Second, we show that the primary features of figure 1—the deterministic production decline and the lack of response to price shocks—hold in subsamples of production from relatively high-volume leases and from wells drilled in-sample. Third, we present results that rule out alternative explanations for the lack of price response.

A.1 Regression analysis for main production and drilling results

Figure 1 indicates graphically that production from previously drilled wells does not respond to oil prices, while figure 2 indicates a strong response of drilling to oil prices. Here, we demonstrate these results more formally via a regression analysis.

As a complement to figure 1, we seek to regress the log of oil production (bbl/d) from wells drilled before 1990 on the logged front month price of oil. We also include as a covariate the percentage difference between the 12-month futures price and the front month price, net of inflation (i.e., the data plotted in the dashed line of figure 3). The oil production data are monthly, and we average the price data to the monthly level. We conduct the analysis in first differences because we cannot reject (using the GLS procedure of Elliott, Rothenberg and Stock 1996) that both log(production) and log(front month price) are unit root processes.[70]

Our main specification is given by:

$$\begin{aligned}\Delta \log(\text{Production}_t) = \alpha + \beta_0 \Delta \log(\text{Price}_t) + \beta_1 \Delta \log(\text{Price}_{t-1}) + \delta_0 \Delta \text{IncreaseRate}_t \\ + \delta_1 \Delta \text{IncreaseRate}_{t-1} + \eta \cdot \text{Time} + \text{error}_t \quad (23)\end{aligned}$$

We include a lagged difference because doing so minimizes the AIC criterion. Removing this lag or adding additional lags does not qualitatively change the results. We include a time trend in our baseline specification to account for the possibility that the pressure-driven production decline curve is hyperbolic rather than exponential, so that $\Delta \log(\text{Production}_t)$ will not be constant over time even if the β and δ coefficients all equal zero (enriching the trend to a polynomial does not qualitatively affect the results). For inference, we use Newey-West with four lags (doing so only slightly increases the estimated standard errors; adding additional lags leaves the estimated errors essentially unchanged).

Column (1) of table 1 presents estimates from the specification in equation (23), while column (2) presents estimates from a specification that does not include the time trend. In both specifications, we find that oil price changes have neither an economically nor statistically significant effect on production from pre-existing wells, consistent with figure 1. In column (1), the sum of the coefficients on the current and lagged difference in log(front month price) yields an insignificant elasticity of oil production with respect to front month price of 0.0009 (with a standard error of 0.034). The sum of the coefficients on the current

[70] For log(production), we obtain a test statistic of -0.643 with the optimal 12 lags, relative to a 10% critical value of -2.548. For log(front month price), we obtain a test statistic of -1.344 with the optimal 11 lags, relative to a 10% critical value of -2.558. Results are similar for all other lags. These results include a time trend in the test, but results are similar when a trend is excluded.

Table 1: First-differenced regressions of Texas production and drilling on oil prices

	Δ log(Production)		Δ log(Drilling)	
VARIABLES	(1)	(2)	(3)	(4)
Δ log(Front-month price)	0.083	0.094	0.372	0.364
	(0.036)	(0.036)	(0.153)	(0.151)
Lagged Δ log(Front-month price)	-0.083	-0.075	0.008	0.003
	(0.036)	(0.037)	(0.188)	(0.187)
2nd lagged Δ log(Front-month price)			0.352	0.346
			(0.165)	(0.162)
Δ Rate of futures price increase	0.0006	0.0008	0.0049	0.0048
	(0.0005)	(0.0005)	(0.0025)	(0.0025)
Lagged Δ rate of futures price increase	-0.0014	-0.0013	-0.0012	-0.0012
	(0.0005)	(0.0005)	(0.0031)	(0.0031)
2nd lagged Δ rate of futures price increase			0.0003	0.0003
			(0.0027)	(0.0027)
Time trend (in years)	0.0005		-0.0004	
	(0.0001)		(0.0009)	
Constant	-0.0112	-0.0071	0.0028	-0.0005
	(0.0012)	(0.0009)	(0.0100)	(0.0050)
N	214	214	213	213
R^2	0.124	0.104	0.110	0.110

Note: Production, drilling, and front-month price data are monthly and the same as in figures 1 and 2. The rate of futures price increase is the percentage difference between the 12-month futures price and the front month price, adjusted for inflation. All prices are in real $2007. Standard errors in parentheses are Newey-West with 4 lags.

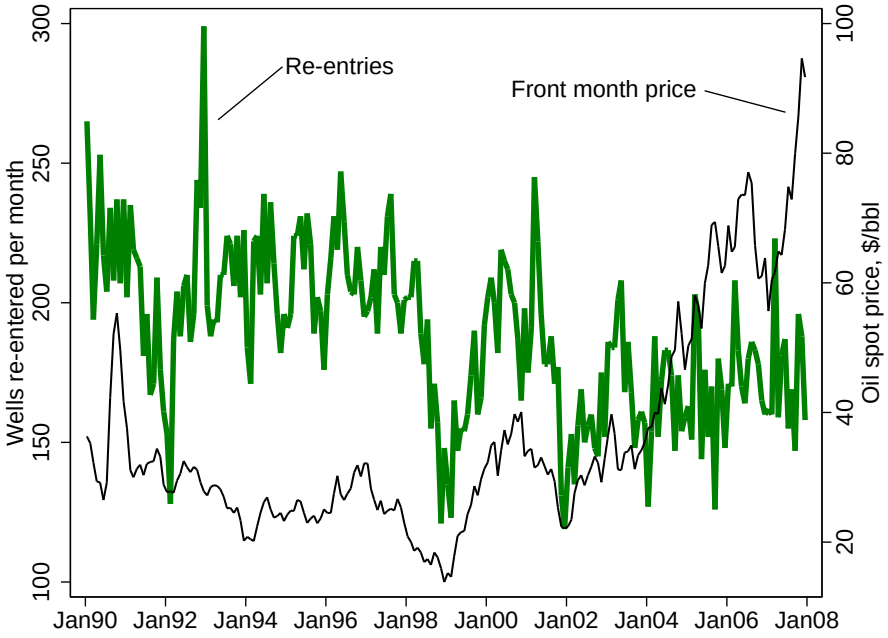
and lagged differences in the rate of futures price increase equals -0.0008, meaning that an increase in the rate of futures price increase of 10 percentage points (about one standard deviation) is associated with only a 0.8% decrease in production. A test against the null hypothesis that this sum equals zero yields a p-value of 0.171.

In columns (3) and (4) of table 1, we re-estimate equation (23) but use the log of wells drilled per month as the dependent variable. We also add an additional lag of the independent variables, minimizing the AIC criterion.[71] Column (3) includes a time trend, while column (4) does not. In either case, the estimates follow what is clear from figure 2: drilling activity responds strongly to changes in oil prices. For column (3), summing the coefficients on the current and lagged front month price differences yields an elasticity of drilling with respect to the front month price of 0.732 (with a standard error of 0.201). The response of drilling to changes in the rate of futures price increase, however, is estimated imprecisely (the sum of the three coefficients equals 0.004, with a standard error of 0.003).

Finally, figure 9 shows a plot of re-entries over time. The use of rigs to re-enter old wells correlates with oil prices, though not as strongly as the drilling of new wells, shown in figure 2.

[71] Intuitively, lags caused by engineering and permitting generate a lag in the response of drilling to price shocks.

Figure 9: Re-entries of existing wells



This figure shows the total number of re-entries across all leases in our dataset. Oil prices are real $2007. See text for details.

A.2 Production decline in high-volume leases and wells drilled in-sample

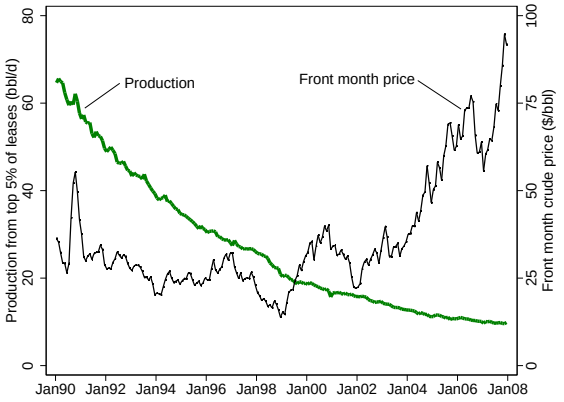
Figure 1 presents average monthly production from all active oil leases in Texas for which there was no rig activity from 1990–2007. Average lease-level production in Texas is quite low, raising the question of whether our empirical results extend to higher-volume fields that might be found elsewhere in the world.

Figure 10 presents production data from subsamples of relatively high-volume leases in Texas. For each lease in the dataset, we obtain its total production by summing its production rate over the entire 1990–2007 sample. We then assign each lease to its appropriate percentile based on total production. The left panel of figure 10 includes leases within the top 5% of total production, and the right panel includes leases in the top 1%. The average top 1% lease produces nearly 200 bbl/d at the start of the sample, far larger than the overall average initial production of about 8 bbl/d shown in figure 1. Nonetheless, in both panels of figure 10, production declines deterministically and exhibits essentially no response to price signals. The production declines are steeper in figure 10 than in figure 1, suggesting either that high-volume leases have relatively high decline rates or that leases typically have decline curves that are hyperbolic rather than exponential.

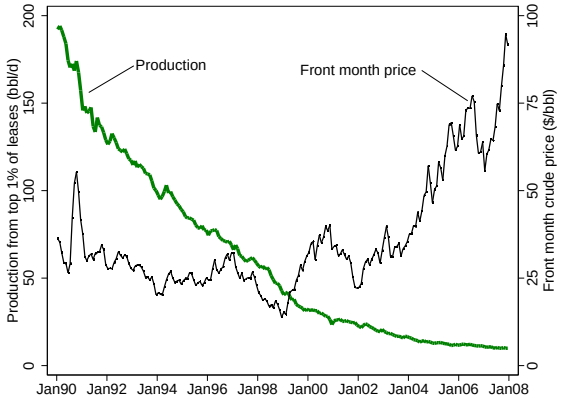
Figure 11 shows production from leases in the 17th vigintile of total production (i.e., leases in the 80th to 85th percentiles of production). The average production rate from these leases is 5.7 bbl/d, so that this vigintile is the closest match to the average production rate (6.3 bbl/d) from leases that were excluded from the sample because they experience

Figure 10: Production from existing wells in high-volume Texas leases

(a) Top 5% of leases



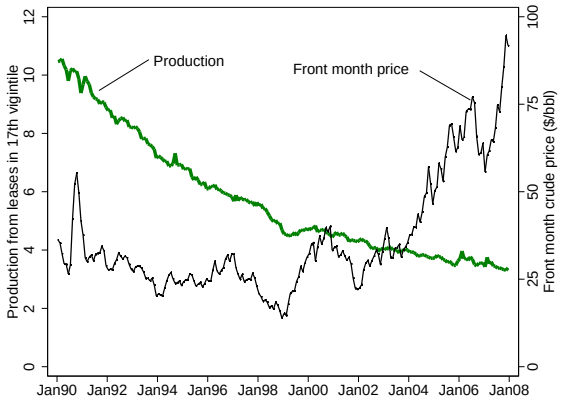
(b) Top 1% of leases



Note: This figure presents crude oil front month prices and daily average oil production from leases on which there was no rig activity (so that all production comes from pre-existing wells). All prices are real $2007. The left panel (a) includes leases that are in the top 5% of total production in the 1990–2007 sample, while the right panel (b) includes leases that are in the top 1%.

rig work between 1990 and 2007. For these leases, production declines deterministically and exhibits essentially no response to price signals.

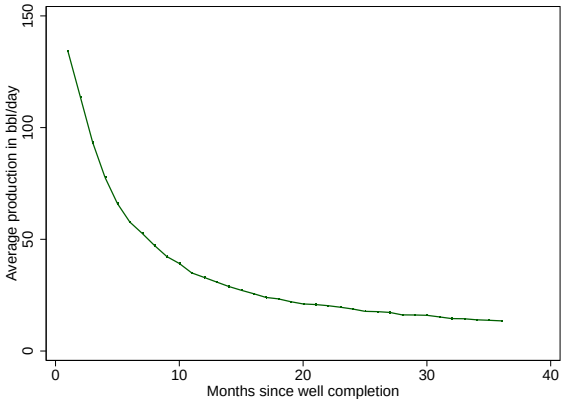
Figure 11: Production from existing wells in the 17th vigintile (80th–85th percentile) of Texas leases



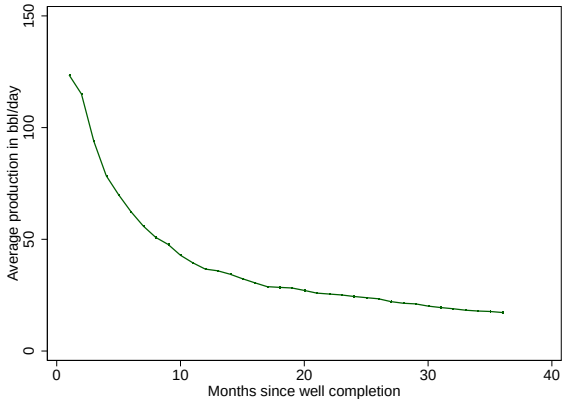
This figure presents crude oil front month prices and daily average oil production from leases on which there was no rig activity (so that all production comes from pre-existing wells). All prices are real $2007. Production data include leases that are in the 17th vigintile (80th–85th percentile) of total production in the 1990–2007 sample.

Figure 12: Production from wells drilled during the sample period

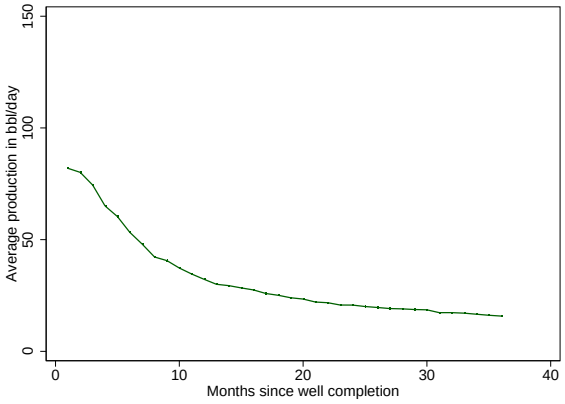
(a) Wells drilled from 1990–1992



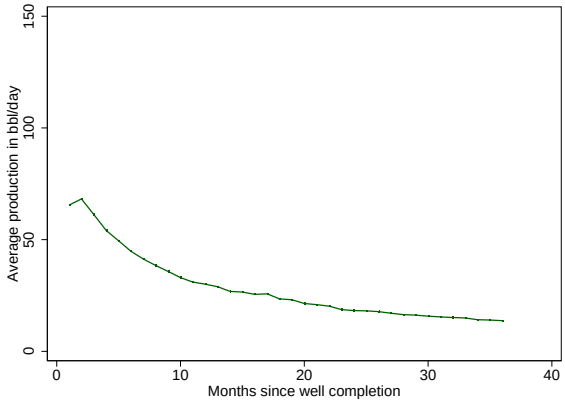
(b) Wells drilled from 1993–1995



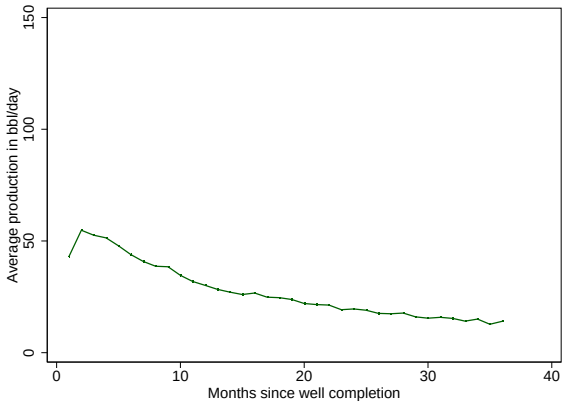
(c) Wells drilled from 1996–1998



(d) Wells drilled from 1999–2001

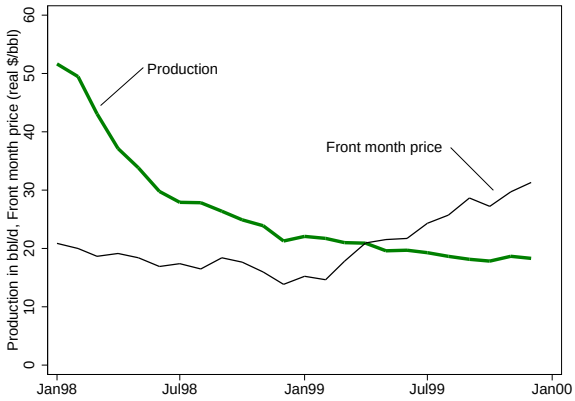


(e) Wells drilled from 2002–2004



Note: Each panel plots average monthly production for wells drilled during the indicated time interval. See text for details.

Figure 13: Production during 1998–1999 from wells drilled in 1997



This figure plots average monthly production for wells drilled in 1997. Oil prices are real $2007. See text for details.

We next examine production from wells drilled during the sample. To undertake this analysis, we first match drilling records, which come from TRRC drilling permits, to the TRRC lease-level production data. This match must be done based on lease names. Because naming conventions vary across the two datasets, we are able to match only 31.6% of drilled oil wells to a lease in the production data.

Because production can only be observed at the lease-level rather than at the well-level, we isolate the sample to drilled wells that are the only producing well on their lease for three years after the well was completed. The remaining sample consists of 5,542 wells drilled between 1990 and 2004 (inclusive). We break these 15 years into five 3-year periods. Each panel of figure 12 then plots, for wells drilled within a particular period, the average production for the first three years of the wells' lives. For each period of drilling, production from drilled wells declines deterministically over time and does not indicate any price responsiveness. To further ensure that we do not observe a price response during 1998–1999, we have isolated wells drilled in 1997, and in figure 13 we plot production from these wells by month against the front month price of oil. We observe no response of production to the 1998–1999 oil price crash.[72]

Production from wells drilled in later periods is lower, on average, than production from wells drilled in early periods, suggesting that firms sensibly drill highly productive wells before drilling less productive wells. Incorporating well-level heterogeneity into the theoretical model is therefore likely to be a fruitful path for future research.

[72]We have conducted a similar analysis for production from re-entered wells, and the resulting plots are similar to what is shown in figures 12 and 13, except that we do not observe a strong difference in overall production rates between wells drilled early and wells drilled late in the sample.

A.3 Ruling out alternative explanations for production's lack of response to price incentives

There are several potential alternative explanations for the lack of price response among existing oil wells that we need to rule out. First, races to oil created by common-pool externalities in oil fields with multiple lease-holders would diminish the incentive that individual lease-holders have to defer production when prices are expected to rise, since much of the deferred production would be extracted by others instead. Figure 14(a) shows, however, that the long, downward trend in oil production manifests both for oil fields with multiple operators and for oil fields with just a single operator.

Second, a condition of many leases is that the lease-holder produce oil; a firm that drops lease-level production to zero may therefore risk losing the lease. Figure 14(b) shows, however, that production from multi-well leases, on which producers may shut in at least some of their wells without risk of losing the lease, does not in general respond to price signals. The only observed response is in 1998–1999, where the data are consistent with shut-ins of wells due to fixed costs, as discussed extensively in appendix C.

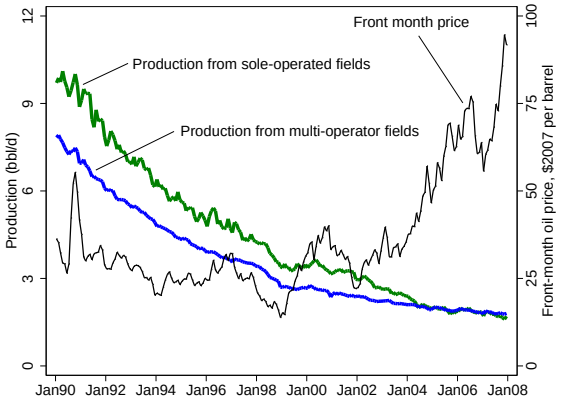
Third, oil production in Texas is subject to maximum allowable production quotas—or "allowables"—as determined by the Texas Railroad Commission. This system dates to the East Texas Oil Boom of the 1930s when a large share of world oil production came from Texas, and races to oil led to overproduction and collapsing world oil prices. Whether originally intended to end the race to oil, or simply to cartelize the Texas oil industry and boost prices, this system persists to this day, and every lease in our data has a monthly allowable, including on fields with just a single operator. One obvious concern is that these maximum production quotas are binding, leading to the lack of price response. Figure 14(c) shows, however, that the average production for leases in our main sample is well below the average allowable production. Thus, the allowables are not binding and therefore cannot explain the lack of price response that we observe in our data.

Fourth and finally, one possible concern is that the decision makers whose behavior we observe in our data could earn profits by delaying production during periods of extreme contango but do not do so because they dynamically optimize incorrectly or because their future price expectations are not aligned with the futures market. Figure 14(d) shows, however, that above-ground storage of crude oil on these leases increased notably during the 1998–1999 period of extreme contango. Lease-holders responded to this price incentive by accumulating inventories above-ground, deferring sales—not extraction—to take advantage of the expected increase in prices, and confirming that they appropriately respond to incentives generated by price expectations aligned with the futures market.[73]

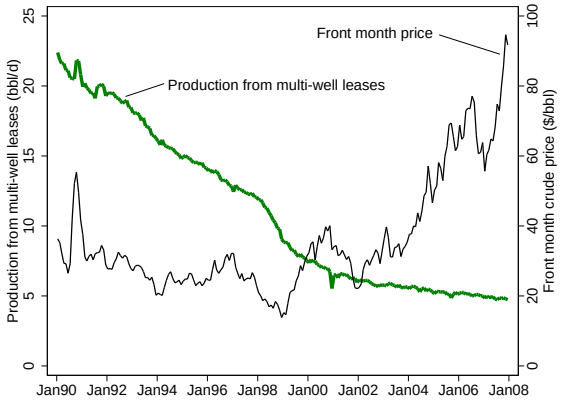
[73] Because changes in on-lease storage volumes are small relative to production rates, oil sales in bbl/d are not substantially different from production throughout the sample.

Figure 14: Graphical evidence ruling out alternative explanations

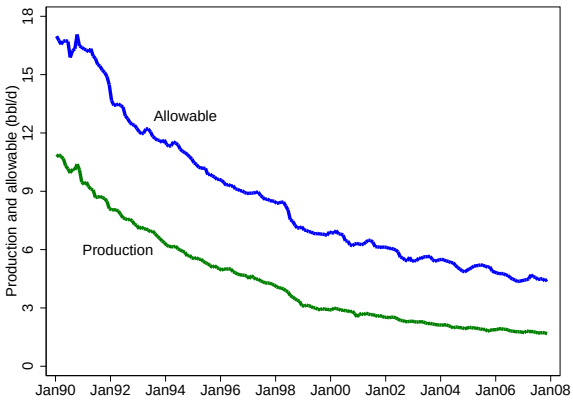
(a) Single vs. multiple-operator fields



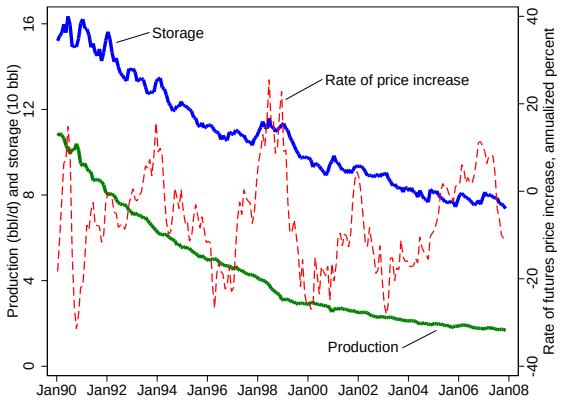
(b) Production on multi-well leases



(c) Actual versus allowable production



(d) Production and above-ground storage



Note: Panel (a) shows average daily production on fields with multiple operators as well as on fields with just a single operator, along with the spot price of crude oil. Panel (b) shows average production for leases with multiple wells in our main sample. Panel (c) shows average actual production as well as average allowable production for leases in our main sample. Panel (d) shows average oil production and the average stock of crude oil stored above-ground on leases in our main sample, along with the annualize rate of futures price increase. See text for details.

B Details for calculations of deferment values discussed in section 4

This appendix describes how we calculate deferment values in section 4. We first describe how we calculate deferment values by using expected future prices directly in place of our model's deterministic future price path. We then describe how we calculate deferment values in a dynamic-stochastic optimization framework.

B.1 Estimation of the value of capacity using only expected future oil prices

Under our assumptions of an annual 10% decline rate and 10% real interest rate, the corresponding monthly decline rate and monthly real rate of interest are then given, respectively, by $\lambda = 1 - (1 - 0.10)^{1/12}$ and $r = (1 + 0.10)^{1/12} - 1$.

When we use futures market prices directly as our measure of future price expectations, the longest futures contract that we typically observe has a maturity of 60 months. We assume that prices more than 60 months in the future were expected at time t to plateau at the level of the 60-month futures price.[74] This assumption naturally follows from the long-run plateauing of the futures curves in the data, as is apparent in figure 3(a).

Given these assumptions, it is straightforward to calculate recursively the value of deferred production at each date t. Since prices were anticipated to plateau after 60 months, it would be optimal to produce at the constraint from that date onward, and so an additional unit of production capacity 60 months hence was anticipated at time t to be worth at $t+60$:

$$\theta(t, 60) = \frac{P(t, 60)}{1 - \delta}, \tag{24}$$

where $\delta = (1 - \lambda)/(1 + r)$ and $P(t, 60)$ is the price expected at t to prevail after 60 months. This is simply the value of inheriting one extra barrel of monthly production capacity in 60 months and producing at capacity forever while earning $P(t, 60 + j) = P(t, 60)$ per barrel for $j > 1$, all discounted back to time $t+60$. Now suppose inductively that the shadow value on capacity $s + 1$ months in the future is anticipated at time t to be $\theta(t, s + 1)$. Then, the anticipated shadow value on capacity s months in the future is given by:

$$\theta(t, s) = \max\left\{P(t, s) + (1 - \lambda)\theta(t, s + 1)/(1 + r), \theta(t, s + 1)/(1 + r)\right\}, \tag{25}$$

where the well owner anticipates in month t choosing s months in the future the more lucrative of the two options: (1) producing at capacity and earning $P(t, s)$ that month and having fraction $1 - \lambda$ of initial capacity remaining in the following month, which is then valued at $\theta(t, s + 1)$; or (2) deferring production until the following month so that the full inherited capacity is saved until the following month, which is again valued at $\theta(t, s + 1)$. Thus, by backward recursion starting 60 months in the future, we can reconstruct the full

[74] Our results are virtually identical when we impose a plateau at the latest-dated futures price beyond 60 months, when available. There are also periodic gaps in the futures data, particularly at longer time horizons. We linearly interpolate prices to fill these gaps.

sequence of shadow values, all the way back to $\theta(t,1)$. Note then that it will be optimal to produce at the constraint at time t whenever $P(t,0)+(1-\lambda)\theta(t,1)/(1+r) > \theta(t,1)/(1+r)$, or equivalently $P(t,0) - \lambda\theta(t,1)/(1+r) > 0$. This condition is the discrete-time analog of necessary condition (7) in the text. So the benefit from deferring one barrel of production at time t is $\lambda\theta(t,1)/(1+r)$, and the cost of deferring it is $P(t,0)$ in month t. The results of this procedure are plotted in figure 4 in the main text (along with the results when we assume a no-change price forecast).

When we use risk-adjusted futures prices to measure firms' price expectations, we must modify our calculations because for a CAPM $\beta \neq 0$, price expectations will not align exactly with futures prices. Rather, price expectations will be higher or lower than the futures price, depending on whether $\beta > 0$ or $\beta < 0$, following equations (18) and (19).

We estimate β using the full sample of monthly crude oil front month prices from April 1983 (when NYMEX trading began) through April 2015, along with the S&P 500 index for the corresponding months. Specifically, we regress monthly front month returns on monthly S&P 500 returns and a constant, computing standard errors via Newey-West with 4 lags. We obtain a point estimate of $\beta = 0.05$, not qualitatively different from zero, with a standard error of 0.24. We have also estimated β using a rolling regression with a six-year window. This series of regressions yields estimates that are lower than our full sample estimate of 0.05 for nearly the entire 1990–2007 period (the largest estimated β is 0.20 in October 2001).

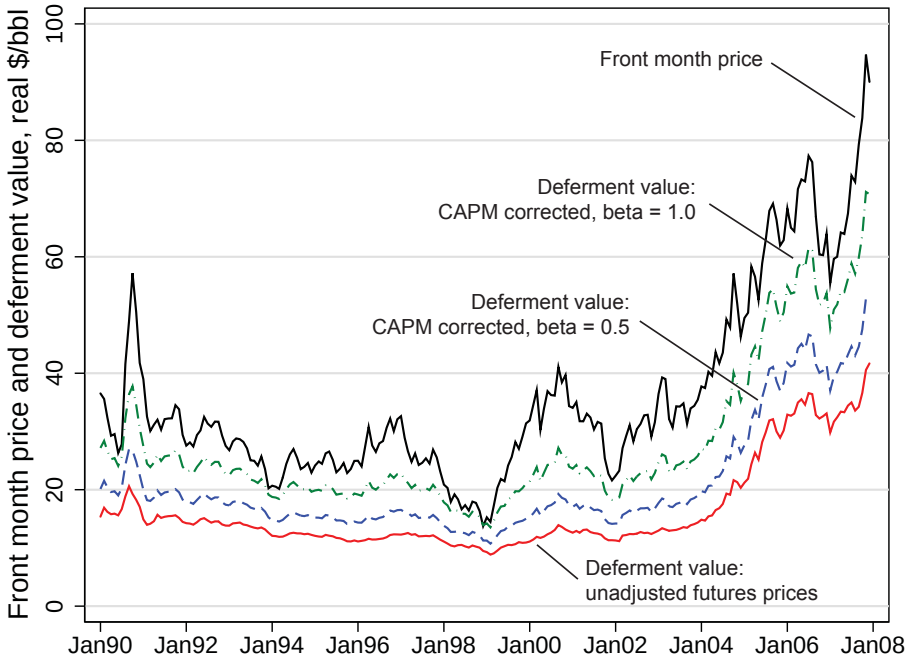
Figure 15 complements figure 4 by comparing the front-month oil price to the value of deferred production ($\lambda\theta(t)$), using values of β of 0.5 and 1.0. $\beta = 0.5$ is just outside the 95% confidence interval of our estimate of $\beta = 0.05$, while $\beta = 1.0$ is well outside of it. To calculate $\theta(t)$ for each value of β, we begin by using equations (18) and (19) in combination with our futures price data (continuing to assume that futures prices plateau at 60 months) to generate expected future prices out to a 30-year (360-month) horizon.[75] Setting $\theta(t,361)$ equal to zero, we then recursively apply equation (25) to generate $\theta(t,1)$ and assess whether production at the constraint is optimal at time t.[76] The results of this procedure for $\beta = 0.0$, $\beta = 0.5$, and $\beta = 1.0$ are plotted in figure 15. Even with $\beta = 0.5$, the spot price of oil always exceeds the value of deferred production, $\lambda\theta(t)$. Figure 15 also shows that a β of approximately 1.0 is required before producers have an incentive to defer production in even one month of the sample (December 1998).

To test the sensitivity of our conclusions to parameter assumptions, we calculate the threshold values of r and λ that would lead the value of deferred production to just equal the spot price during the 1998–1999 episode. For simplicity, we use unadjusted futures prices ($\beta = 0$) as the measure of firms' price expectations. We find that if either firms' discount rate were as low as 4% annually or the production decline rate were as high as 30% annually, then it would have been optimal to defer production briefly during the 1998–1999 episode. Similarly, we calculate a threshold rate of increase in anticipated oil prices beyond the 60-

[75] For annualized market and riskless interest rates, we use values from Damodaran (2015b): 11.53% and 3.53% (both nominal). The value 11.53% is the average S&P 500 return from 1928–2014, and 3.53% is the average three-month treasury bill rate from 1928–2014.

[76] To verify that a 30-year horizon is sufficiently long to yield a precise solution for $\theta(t,1)$, we apply this procedure to unadjusted futures prices and find that the calculated value for $\theta(t,1)$ is generally within 3 cents/bbl of the value calculated when an infinite horizon is assumed (i.e., when equation (24) is used to calculate $\theta(t,60)$).

Figure 15: Crude oil spot price vs. value of deferred production at $\beta = 0.5$ and $\beta = 1.0$



Note: This figure shows the crude oil front month ("spot") price and the value of one barrel of deferred production ($\lambda\theta(t)$) in each month, all in real 2007 dollars. The time series showing the value of deferred production uses price expectations based on unadjusted futures prices and on risk adjusted futures prices using CAPM β's of 0.5 and 1.0. See text for details.

month limit of our futures data. We find that if anticipated prices were to rise at a rate of 11% annually from the 60-month futures price, rather than plateau at that level, then it would again have been optimal to defer production briefly during the 1998–1999 episode.

B.2 Estimation of the value of capacity allowing for stochastic oil prices

Let S_t denote the state vector the agent observes at t, which includes the spot price P_t and statistics for future prices but may also include other variables he cannot influence. Let $Pr(S_{t+1}|S_t)$ be a stationary Markov transition matrix for S_t. Consider our infinite-horizon, stationary programming problem in the presence of uncertainty.[77] It can be verified that it is optimal for a risk-neutral agent to produce at capacity if $P_t \geq \lambda\theta(S_t)/(1+r)$, where the function $\theta(S_t)$ denotes the value currently expected to prevail next month for each additional unit of capacity brought into that period. Intuitively, for each additional unit the agent produces in the current period, he receives P_t more in revenue but enters the next period with λ less capacity than he would if he did not produce that unit. Given S_t, he

[77] $V_t(K_t, S_t) = \max_{F_t \in [0,K_t]} P_t F_t + \frac{1}{1+r} \sum_{\text{all } S_{t+1}} V_{t+1}(K_t - \lambda F_t, S_{t+1}) Pr(S_{t+1}|S_t)$.

currently expects a unit of capacity next period to be worth $\theta(S_t)$ in next period dollars. It is optimal to produce one unit more if $P_t \geq \lambda\theta(S_t)/(1+r)$. Indeed, since neither side of this inequality depends on the size of the sale, he should produce at capacity whenever this inequality holds. We interpret $\lambda\theta(S_t)/(1+r)$ as the deferment value in a stochastic setting.

The value $\theta(S_t)$ must satisfy the following recursive equation:

$$\theta(S_t) = \sum_{\text{all } S_{t+1}} \max\left\{P_{t+1} + (1-\lambda)\theta(S_{t+1})/(1+r), \theta(S_{t+1})/(1+r)\right\} Pr(S_{t+1}|S_t). \quad (26)$$

This functional equation defines the function $\theta(\cdot)$ uniquely, since it satisfies Blackwell's sufficiency condition for a contraction mapping. Intuitively, given the current state S_t, the agent anticipates that, for each state that may occur next period, he will utilize an additional unit of capacity in the optimal way from that period onward and will therefore receive in expectation $\theta(S_t)$.

Our first task is to characterize and estimate each month's state vector, which includes statistics for crude oil spot and futures prices. To limit the dimensionality of our state space, we assume that each month's futures curve can be characterized by two state variables: (1) the crude oil spot price (P_t) and (2) a long-run steady-state futures price (P_t^*). We then assume that futures prices at t follow an exponential curve given by equation (27):

$$F_{ts} = \pi^s P_t + (1-\pi^s)P_t^*, \quad (27)$$

where F_{ts} denotes the futures price at time t for delivery at time $t+s$ (i.e., the s-month futures price at time t), and $\pi \in (0,1)$ is a constant term that reflects the proportional rate of convergence to the steady state. Thus, the futures price at time $t+s$ is a weighted average of the spot price and long-run steady-state price at time t, with the weight on the spot price given by the decaying value π^s. This functional form is consistent with the observation that futures curves in our sample typically rise or fall asymptotically toward some long-run plateau.

While the P_t are directly observed, we must estimate the P_t^* and π using our futures data (aggregated to the monthly level, as always). The empirical analog of equation (27) is given by the following equation:

$$F_{ts} = \pi^s P_t + (1-\pi^s)P_t^* + \epsilon_{ts}, \quad (28)$$

where both F_{ts} and P_t are given by our futures data, and ϵ_{ts} is an error term. We jointly estimate π and P_t^* using the following two-step Nonlinear Least Squares (NLS) estimator:

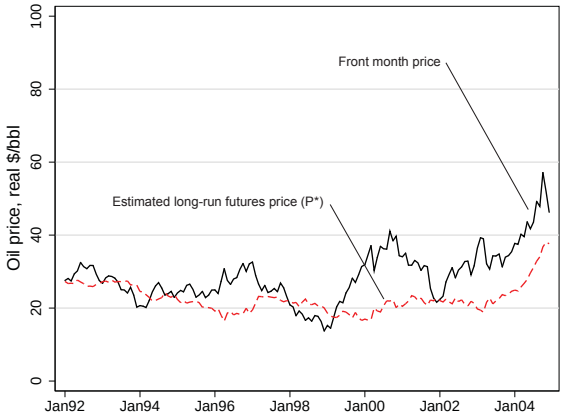
1. Provisionally guess $\tilde{\pi}$.

2. Estimate the P_t^* values in equation (28) via an OLS regression of $F_{ts} - \tilde{\pi}^s P_t$ on the set of interactions $(1-\tilde{\pi}^s) \times D_t$ for all t, where the D_t is a time-t dummy variable. Label the corresponding coefficient estimates $P_t^*(\tilde{\pi})$ for all t.[78]

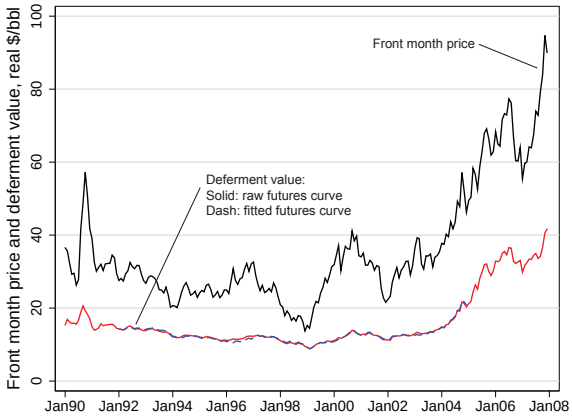
[78] In practice, we estimate P_t^* via a Weighted Least Squares (WLS) regression of $(F_{ts} - \tilde{\pi}^s P_t)/(1-\tilde{\pi}^s)$ on a set of time-t fixed effects, with analytical weights given by $(1-\tilde{\pi}^s)^2$, which is numerically equivalent but computationally faster.

Figure 16: Crude oil spot prices, estimated steady-state futures prices, and deferment values calculated using raw and fitted futures data

(a) Spot price and estimated steady-state futures price



(b) Spot price and deferment values



Note: In panel (a) the solid black line shows crude oil front month ("spot") prices, and the dashed line shows the steady-state futures price (P_t^*) estimated using the procedure described in the text. In panel (b) the solid black line shows the spot price, the solid red line shows the estimated deferment value calculated using our raw futures data and the methods described in section B.1, and the dashed blue line shows the estimated deferment value calculated using our fitted futures data using the identical method. See text for details.

3. Construct each observation's residual based on equation (28), conditional on the provisional guess of $\tilde{\pi}$: $\epsilon_{ts}(\tilde{\pi}) = F_{ts} - \tilde{\pi}^s P_t - (1 - \tilde{\pi}^s) P_t^*(\tilde{\pi})$.

4. Estimate π via NLS: $\pi^{NLS} = \text{argmin}_{\tilde{\pi}} SSR(\tilde{\pi})$, where $SSR(\tilde{\pi}) = \sum_{ts} \epsilon_{ts}(\tilde{\pi})^2$.

5. Return to step 2 to estimate final expected steady-state prices conditional on the NLS estimate π^{NLS}: $P_t^*(\pi^{NLS})$ for all t.

We fit this parametric model to our futures data for the 1992–2004 time period, during which the behavior of spot and futures prices was relatively stable. Thus, we also focus our empirical tests on the 1992–2004 time period, which overlaps the key 1998–1999 period of severe contango. Using earlier or later data tends to yield an estimated π that is too large, leading to a poor fit of the futures data and estimated P_t^* values that vary substantially and that can even be negative. We estimate $\pi^{NLS} = 0.954$ for the 1992–2004 period (this value is a monthly decay rate). Figure 16(a) plots our corresponding estimates for P_t^* along with the spot price. Not surprisingly, our estimated P_t^* values tend to exceed the spot price during periods of severe contango in the futures market. Figure 16(b) shows the deferment value calculated using our fitted futures data using the method described in section B.1 above, along with the deferment value calculated using our raw futures data using the same method. This figure demonstrates that our raw and fitted futures data yield very similar estimates

for the deferment values during the 1992–2004 period.

Our second task is to estimate a first-order Markov process for our state vector. Our main analysis takes futures prices directly as expected future prices. Thus, our parametric model for the futures curve in equation (27) implies the following price expectations:

$$\begin{aligned} E_t[P_{t+1}] &= \pi P_t + (1-\pi)P_t^* \\ E_t[P_{t+1}^*] &= P_t^*. \end{aligned} \tag{29}$$

Denote by $\epsilon_P = \ln(P_{t+1}) - \ln(E_t[P_{t+1}])$ and $\epsilon_{P^*} = \ln(P_{t+1}^*) - \ln(E_t[P_{t+1}^*])$ the monthly log-innovations to the spot price (P_t) and long-run steady-state futures price (P_t^*). We assume that these log-innovations are jointly distributed iid normal:

$$\begin{bmatrix} \epsilon_P \\ \epsilon_{P^*} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\epsilon_P} \\ \mu_{\epsilon_{P^*}} \end{bmatrix}, \begin{bmatrix} \sigma^2_{\epsilon_P} & \rho\sigma_{\epsilon_P}\sigma_{\epsilon_{P^*}} \\ \rho\sigma_{\epsilon_P}\sigma_{\epsilon_{P^*}} & \sigma^2_{\epsilon_{P^*}} \end{bmatrix} \right), \tag{30}$$

where μ_P and μ_{P^*} denote the means of ϵ_P and ϵ_P^*, σ_P^2 and $\sigma_{P^*}^2$ denote their variances (and σ_P and σ_{P^*} their standard deviations), and ρ denotes the correlation coefficient between these variables. Note that equation (29) and the iid assumption together imply that P^* follows a random walk. We impose the mean restrictions that $\mu_{\epsilon_P} = -\sigma^2_{\epsilon_P}/2$ and $\mu_{\epsilon_{P^*}} = -\sigma^2_{\epsilon_{P^*}}/2$, which, given our assumption of normality for log-innovations, ensures that price expectations in levels conform to the parametric model in equation (29) above. Given these distributional assumptions and mean restrictions, we then estimate parameters $\sigma_P^2 = 0.00612$, $\sigma_{P^*}^2 = 0.00202$, and $\rho = -0.434$ via maximum likelihood using our observed P_t and estimated P_t^* data; the corresponding annualized volatilities are 35% for P_t and 18% for P_t^*.

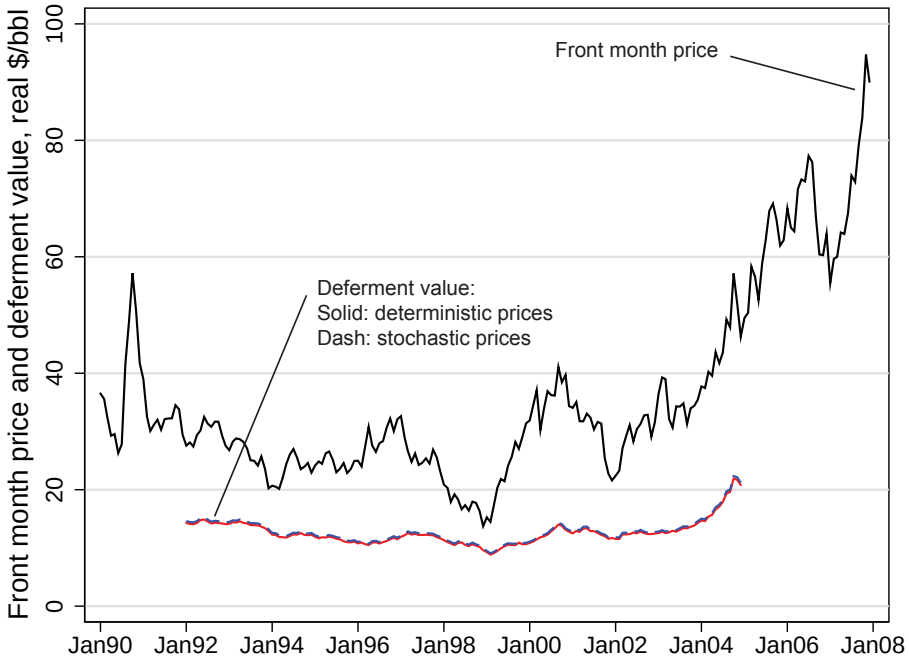
Our third task is to computationally solve the well owner's maximization problem using value function iteration. Our state space consists of 40 steps in each of our two dimensions (i.e., 1600 discrete states), ranging in each dimension from one-third times the lowest price observed in our sample to three times the highest price.[79] We use our maximum likelihood parameter estimates for the first-order Markov process above to calculate the state transition matrix associated with this state space; i.e., the probability, given an initial state, of moving to any of the 1600 discrete states. The numerical solution to the optimization problem in equation (26) yields a policy function for whether to produce or not, as well as the value function, at all possible discrete states. We can then calculate the deferment value associated with each of our discrete states: $\lambda\theta(S)/(1+r)$.

Our fourth and final task is to calculate the time series of deferment values during our 1992–2004 sample period. As discussed above, each month's state is characterized by our observed P_t and estimated P_t^* values. We calculate the corresponding time series of deferment values via linear interpolation across the four discrete states that bound each month's two-dimensional state vector.

Figure 17 presents the results of our analysis. As indicated in the figure, the deferment value based on our stochastic model lies uniformly above the deferment value based on our deterministic model, which to ensure an apples-to-apples comparison takes the fitted futures curve directly as its measure of future prices. However, the difference between these two

[79] We found that extending the state space to 50 steps in each dimension, ranging from one-quarter to four times the lowest and highest prices observed in our sample, had a trivial impact on our estimates.

Figure 17: Crude oil spot price vs. value of deferred production



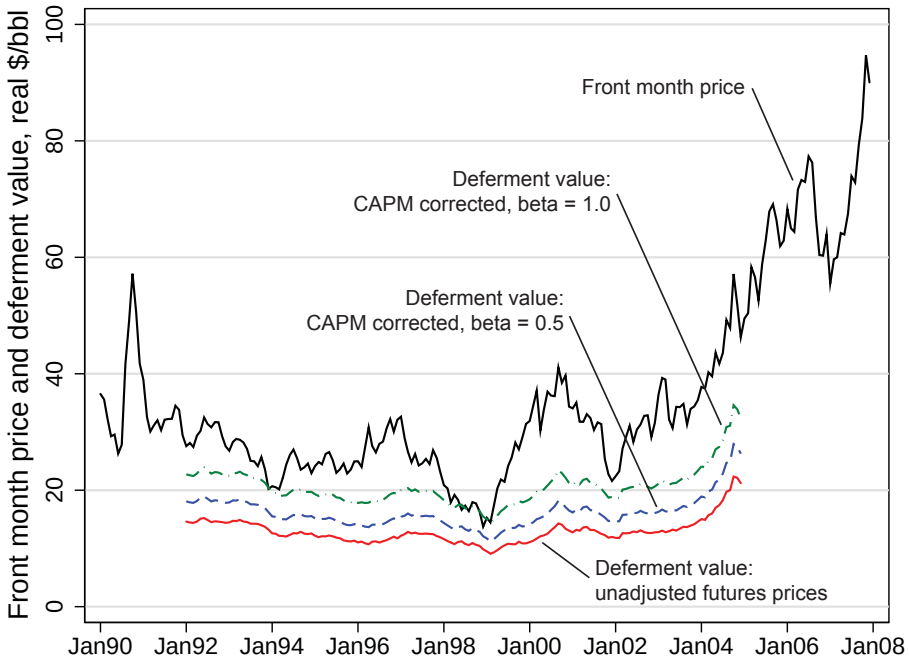
Note: This figure shows the crude oil front month ("spot") price and the value of one barrel of deferred production in each month based on our stochastic model (dashed blue line) and based on using our fitted parametric futures curve directly in our deterministic model (solid red line). See text for details.

deferment values is extremely small—always less than $1—such that the deferment value based on our stochastic model still lies well below the spot price in every month. Thus, in practice, the option value associated with stochastic prices is quite small.

Parallel to our analysis in the main text, we also consider how adjusting futures prices for non-diversifiable risk in a CAPM framework would affect our conclusions. Our procedure is nearly identical to that outlined above. We begin by fitting the same parametric function to our futures curve, yielding identical estimates of the π and P_t^* values. However, equation (29) relating the futures curve to price expectations no longer holds, which requires a minor modification to our procedure. We adjust $E_t[P_{t+1}]$ to account for a CAPM $\beta > 0$, which is straightforward using equations (18) and (19). Adjusting $E_t[P_{t+1}^*]$ is trickier: we use the the fact that $E_t[P_{t+2}] = E_t[E_{t+1}[P_{t+2}]]$, along with the shape of the parametric futures curve and structure of the CAPM adjustment, to back out the value of $E_t[P_{t+1}^*]$ that yields the correct value for $E_t[P_{t+2}]$. We then return to the procedure outlined above, calculating log-innovations to P_t and P_t^*, estimating the corresponding covariance matrix of their joint normal distribution via maximum likelihood, and so forth, eventually yielding a time series of monthly deferment values using the procedures described above.

Figure 18 presents the results using CAPM beta values of $\beta = 0.5$ and $\beta = 1.0$, along with the results based on unadjusted futures prices ($\beta = 0$) for comparison. As in the main text,

Figure 18: Crude oil spot prices and deferment values using futures prices that have been adjusted for non-diversifiable risk in a CAPM framework



Note: This figure shows the crude oil front month ("spot") price and the value of one barrel of deferred production in each month based on our stochastic model using unadjusted futures prices (solid red line) and based on our stochastic model using futures prices that have been adjusted for non-diversifiable risk in a CAPM framework with CAPM betas of 0.5 and 1.0 (dashed blue and green lines). See text for details.

using unadjusted futures prices yields nearly identical results as when we assume $\beta = 0.05$, which is the value we estimate based on a regression using our full sample.[80] As in the main text, large values for the CAPM beta are required—not as high in our deterministic model, but greater than $\beta = 0.5$—for deferring production ever to be optimal during our sample period.

[80] We do not attempt to replicate our analysis using time-varying β values estimated via recursive or rolling regressions, since doing so would require adding yet another state variable to our analysis.

C Extending the model to include fixed costs

C.1 Evidence of shut-ins of low-volume wells in 1998–1999

In section 2.2, we noted a slight deviation of production from its long-run trend during the 1998–1999 period in which the spot price fell below $20/bbl and the rate of futures price increase exceeded 10% (and sometimes 20%) on an annual basis. In particular, it appears that production accelerated its decline rate in 1998 while prices were falling, leveled off in 1999 while prices were rising, and then resumed its usual decline in 2000.

To assess whether this deviation is real and what mechanism lay behind it, we study whether it arose from wells being shut in or from changes in production on the intensive margin. We first isolate the sample to leases that had no more than one flowing well over 1994–2004,[81] so that observed lease-level production during this time can be interpreted as well-level production. We then split this sample into two groups: wells that were never shut in over 1994–2004 and "intermittent" wells that were shut in for at least one calendar month. Panel (a) of figure 19 plots the time series of production from these two samples. This figure makes clear that the 1998 deviation from trend was driven entirely by wells that sometimes have zero production. For wells that always produced, there is no adjustment on the intensive margin, even though firms are able to adjust their pumping rate without shutting in (see Rao 2010).

To explore further, panel (b) of figure 19 indicates, for our sample of one-well leases from 1994–2004, the share of wells that were shut in each month. Overall, there was a steady increase in shut-ins over time. When prices fell in 1998, an unusually large number of wells were shut in, leading to the steepening of the overall production rate decline shown in panel (a) of figure 19. Then, when prices recovered during 1999, the share of shut-in wells flattened, temporarily slowing the production decline.

Finally, we study the selection of which wells were shut in during 1998–1999. We first isolate our sample to wells that produced a non-zero volume of oil every month during 1997, leaving us with 4,605 wells. We then identify whether each of these wells (a) was shut in for every month in 1999; (b) produced oil for at least one month in 1999; or (c) produced oil every month in 1999. We then run a simple regression of each well's average daily production rate in 1997 on dummy variables for the well's 1999 status. We find that wells in each of the three categories above produced, on average, 2.40 bbl/d, 3.53 bbl/d, and 3.92 bbl/d respectively in 1997.[82] Thus, the wells that shut in during 1998–1999 were likely to have been particularly low-volume, marginal wells.

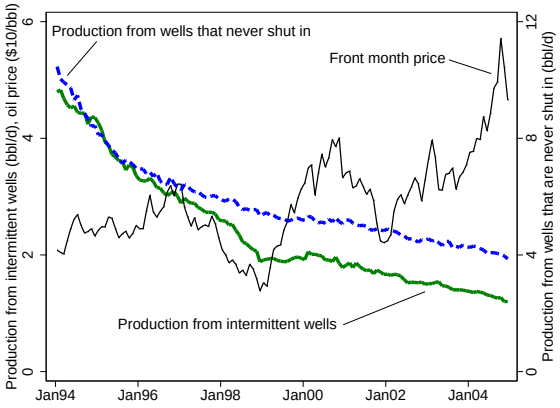
In sections C.2 and C.3 below, we show that these empirical results on shut-ins of marginal wells are consistent with a natural extension of our model that incorporates fixed operating costs.

[81] We use a shorter sample window to increase the number of qualifying leases in the sample and to improve the visualization of the 1998–1999 period in figure 19.

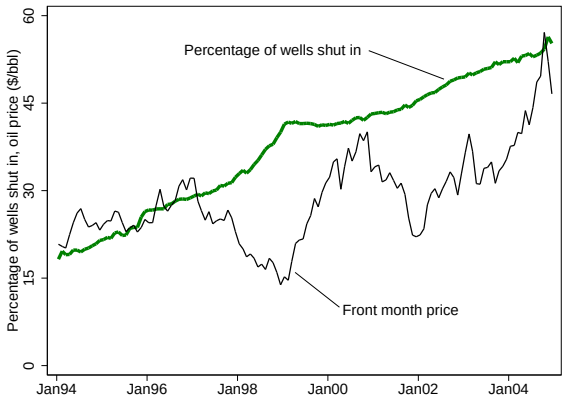
[82] The regression is run with category (a) omitted. Standard errors for the category (b) and (c) dummy variables are 0.47 and 0.25 respectively, so that each of these categories is statistically significantly different from category (a) at the 5% level.

Figure 19: Analysis of well shut-ins

(a) Intermittent wells versus wells never shut in



(b) Percentage of wells that were shut in



Note: Panel (a) shows production from wells that shut in at least once during 1994–2004 versus production from wells that never shut in during 1994–2004. Panel (b) shows the share of wells that were shut in each month. All production data come from leases that had no more than one productive well and never experienced a rig intervention over 1994–2004. Oil prices are real $2007. See text for details.

C.2 Implications of fixed costs when the oil price is exogenous and constant

While our model is in general not tractable in the presence of non-zero fixed costs, progress can be made for the case when firms face an exogenous, constant price. As in section 5, denote this price by $\bar{P}$.

When a well is drilled at t, the cost of renting the rig is $d(a(t))$ and the well's flow is initially X. It then declines exponentially. Hence the flow of revenue at time $y \geq t$ is $\bar{P}Xe^{-\lambda(y-t)}$. Let A denote the avoidable fixed costs. Once the revenue flow declines to A, production should cease. Denote the time when production ceases as $t + t^*$. Then

$$\bar{P}Xe^{-\lambda t^*} = A. \tag{31}$$

That is, when the revenue flow no longer covers the avoidable fixed cost, the well is closed.

At time t drilling another well requires renting a rig at a cost $d(a(t))$. At time $y \geq t$, the flow from the well is $Xe^{-\lambda(y-t)}$. Since the price is assumed to remain $\bar{P}$, at time $y \in [t, t+t^*]$ the well will generate $\bar{P}Xe^{-\lambda(y-t)}e^{-r(y-t)}$ in revenue discounted to time t. The value of drilling a unit of capacity at time t is therefore $\int_{y=t}^{t+t^*}[\bar{P}Xe^{-\lambda(y-t)} - A]e^{-r(y-t)}dy = \frac{\bar{P}X}{r+\lambda}(1 - e^{-(r+\lambda)t^*}) - \frac{A}{r}(1 - e^{-rt^*})$. This anticipated revenue flow is smaller than in our original model, where the formula reduced to $\frac{\bar{P}X}{r+\lambda}$ (since $A = 0$ and $t^* = \infty$). The fixed cost lowers the net revenue flow from a well and shortens the time span that the well remains in service. As a result, wells are less attractive and drilling proceeds at a slower rate.

Empirically, the impact of fixed costs on the rate of drilling is likely to be minor. In

our data, the average drilled well for which we observe production (that is, the wells whose production is plotted in figure 12) initially produces about 100 bbl/d. Suppose the oil price is \$20/bbl and that r and λ are both given by 0.1 (annually). Then, for fixed costs of \$20/day (the value used in the simulations in section C.3 below), $t^* = 46.1$ years and the value of an undrilled well decreases by only 2.0%. For a fixed cost of \$40/day, $t^* = 39.1$ years and the value of an undrilled well decreases by only 4.0%.[83]

With fixed costs, the new equilibrium system is:

$$\frac{\bar{P}X}{r+\lambda}(1-e^{-(r+\lambda)t^*}) - \frac{A}{r}(1-e^{-rt^*}) - d(0) = \gamma_0 e^{rT} \tag{32}$$

$$\frac{\bar{P}X}{r+\lambda}(1-e^{-(r+\lambda)t^*}) - \frac{A}{r}(1-e^{-rt^*}) - d(a(t)) = \gamma_0 e^{rt} \tag{33}$$

$$\int_{t=0}^{T} a(t)dt - R_0 = 0, \tag{34}$$

where t^* is defined in equation (31).

In this case, each producing well operates at capacity, but $\dot{K} = \dot{F}$ has an extra term:

$$\dot{F} = a(t)X - \lambda F(t) - a(t-t^*)Xe^{-\lambda t^*}. \tag{35}$$

The last term reflects the fact that every well turns off when it reaches age t^*, at which point the flow of $Xe^{-\lambda t^*} = A/\bar{P}$ disappears. Wells turning off at t must have turned on t^* periods earlier, at time $t - t^*$. The rate of shut-in at time t therefore depends on the drilling rate at time $t - t^*$, denoted $a(t - t^*)$, which is dictated by the indifference condition (33). Overall, the share of wells that are shut in will increase steadily over time, and at time $T + t^*$, all wells will be shut in so that the total production rate is zero.

Finally, if we shock this extended model with an unanticipated permanent cut in the constant price, the date when each well's production ceases, determined in equation (31), will shorten ($\frac{dt^*}{dP} = \frac{1}{\lambda P} > 0$). Denote the old date as t^* and the new date as $t^{**} < t^*$. Then, upon impact of the price cut, the mass of wells older than t^{**} (all of which will be younger than t^*) will immediately shut down. This shut-down is anticipated to be permanent. Of course, if there is an unanticipated increase in the price, these wells could be returned to service. In section C.3 below, we capture these dynamics using actual data on spot prices and price expectations and a numerical model that relaxes the no-change assumption used here to derive analytic results.

C.3 Extending our analysis of section 4 to include fixed costs

In section 4, we showed that, absent fixed costs and given reasonable estimates of expected future oil prices, it was optimal for firms to maintain production at its capacity constraint

[83] Anecdotally, fixed costs between \$20 and \$40 per day—especially for the 1998 to 1999 period we focus on—seem reasonable in light of recent (and therefore inflated) accounts in the press. Financial Times (2014) cites two operators claiming stripper well operating costs of "thousands of dollars a year" or "in the 'high \$30s' per barrel." Denver Post (2008) reports costs between \$10 and \$25 per barrel. Both Reuters (2015) and Wood Mackenzie (2015) report costs between \$20 and \$50 per barrel.

throughout the 1990–2007 sample, including during 1998–1999. In this section, we examine whether a fixed production cost that must be paid whenever production is non-zero can rationalize the shut-ins of marginal wells that we observe during 1998–1999.

Our model is not tractable in the presence of time-varying prices and a non-zero fixed cost, which we denote by A. We therefore turn to a numerical simulation, with discrete time measured in months. For simplicity, we conduct our simulation assuming that future oil prices are deterministic. This approach ignores the option value that is generated when firms with fixed costs are faced with stochastic oil prices and will therefore understate firms' incentives to shut in low-volume wells for given values of fixed costs and expected future prices.

Consider a firm at time t that owns a well with a capacity of $K(t)$, with price expectations for future periods given by futures prices. As discussed in appendix B.1, our baseline calculations assume that price expectations are no-change beyond 60 months. Thus, for all times $t+T$, $T \geq 60$, the firm's decision rule is simple and given by the empirical analog to equation (31): produce if $K(t+T) \geq A/E[P_{t+60}]$, and shut in otherwise. Thus, given $E[P_{t+60}]$, there is a critical K^*_{60} for which the firm is indifferent about shutting in. If $K(t+60)$ exceeds $K^*_{t,60}$, the firm will produce at the constraint at $t+60$, and it will continue to produce at the constraint until the production decline (at rate λ) leaves it with a capacity less than $K^*_{t,60}$. At this time, which we denote $t+T_f$, the well will shut in forever.

Define $\delta \equiv 1/(1+r)$ (note that this is a change in notation from appendix B.1). Then, given some 60-period ahead capacity $K(t+60) \geq K^*_{t,60}$, the value of that capacity at time $t+60$ is given by

$$V_{t+60}(K(t+60)) = \frac{K(t+60)E[P_{t+60}](1-\delta^{T_f+1}(1-\lambda)^{T_f+1})}{1-\delta(1-\lambda)} - \frac{A(1-\delta^{T_f+1})}{1-\delta}, \qquad (36)$$

while the value of capacity $K(t+60) < K^*_{t,60}$ is zero.

With $V_{t+60}(K(t+60))$ calculated for a large number of discrete capacity states, our numerical simulation can then work backwards from time $t+60$ to time t, assessing whether it is optimal to produce or shut in at each future time for each capacity state. Specifically, it will be optimal to produce at time $t+T$ given capacity K_{t+T} if the following inequality holds:

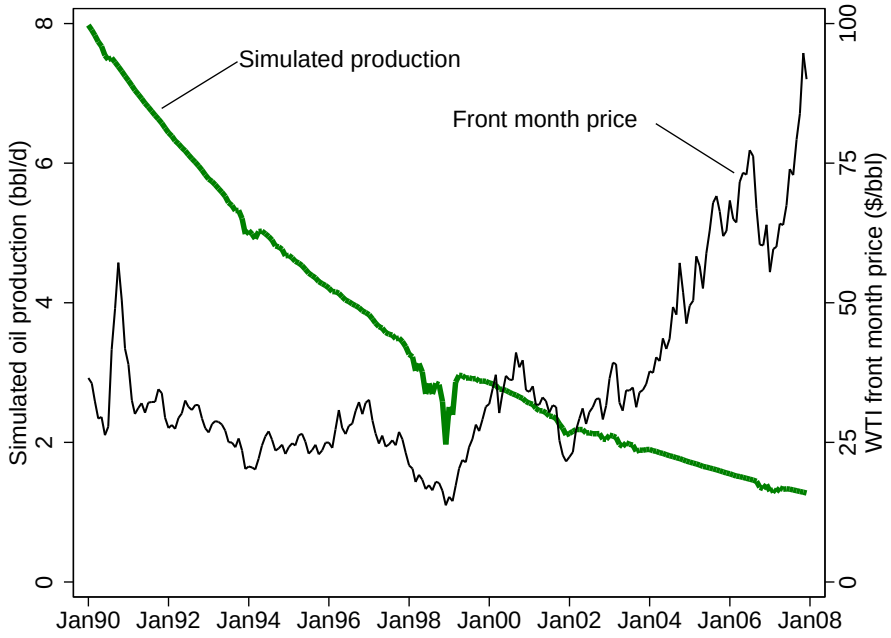
$$E[P_{t+T}]K_{t+T} - A + \delta V_{t+T+1}((1-\lambda)K(t+T)) \geq \delta V_{t+T+1}(K(t+T)). \qquad (37)$$

We measure $E[P_{t+T}]$ using futures prices for each t and T, as was the case in our original calculation of production incentives that generated figure 4. Our recursion procedure produces, for each month t of the sample, a critical capacity K^*_t such that it is optimal to produce a well if and only if its capacity exceeds K^*_t.

We then simulate production from a continuum of wells that initially (in January 1990) had capacities distributed exponentially with a mean of 8 bbl/d (thus, the initial average production rate roughly matches figure 1). Using a fixed cost of \$20 per day,[84] we simulate production from this distribution of wells and plot the average production rate in figure 20.

[84]For larger fixed costs, the simulation begins to predict short-lived but noticable reductions in production during the brief price drops near January 1994 and September 2001.

Figure 20: Simulated production from model with fixed costs



Note: This figure shows simulated production from a set of wells with initial capacities exponentially distributed with a mean of 8 bbl/d. The wells have a fixed operating cost of $15 per day. Oil prices are real $2007. See text for details.

The simulated production shows a distinct dip in 1998–1999, driven by the shut-in and restart of marginal wells. The critical capacity level K_t^* during this period is generally between 1.0 and 2.5 bbl/d, though it briefly spikes up to 4.0 bbl/d in December 1998. The duration of time over which wells are simulated to shut in and restart is narrower than what is observed empirically. This difference is likely driven by (1) non-modeled heterogeneity in fixed costs, (2) the presence of marginal costs that, while small, are not exactly zero, (3) costs associated with re-starting wells, and (4) contractual constraints that will cause some lease-holders to lose their leases if they shut in their wells for a sufficiently long time. Overall, however, our simulation shows that an extension of our model that includes fixed costs is capable of qualitatively rationalizing the shut-ins of marginal wells that we observe in the data.

D Standard Hotelling result

In this section, we briefly restate the standard Hotelling result before illustrating the conditions under which the path of Hotelling's planner and that of our planner coincide.

Assume that oil flow $F(t)$ at time t generates instantaneous utility flow of $U(F(t))$, with $U(0) = 0$, $U'(\cdot) > 0$, and $U''(\cdot) < 0$. Assume that an initial stock of Q_0 units of oil can be extracted at rate $F(t) \geq 0$, which is under the complete control of the oil extractor, and that the cost of this extraction is $cF(t)$, where $c \in [0, U'(0))$ is the constant marginal cost of extraction. Assume the social planner discounts utility and costs continuously at exogenous rate r and, if wealth maximizing agents are involved, they also discount profit flows at rate r. To maximize the discounted utility, the planner chooses $F(t)$ so that marginal utility less marginal cost of extraction grows exponentially at the rate of interest whenever oil flow is strictly positive:

$$F(t) \geq 0,\ U'(F(t)) - c - \gamma_0 e^{rt} \leq 0,\ \text{with complementary slackness (c.s.)}, \tag{38}$$

where γ_0 is a multiplier. Thus, quantity flows according to: $F(t) = U'^{-1}\left(\gamma_0 e^{rt} + c\right)$, where U'^{-1} is the inverse of the first derivative of $U(\cdot)$. In addition, the resource stock must be completely extracted either in finite time or asymptotically: $\int_0^\infty F(t)dt = Q_0$, which uniquely determines γ_0 and therefore the time path of extraction and marginal utilities.

If we assume that marginal utility is unbounded at zero ($U'(0) = \infty$), then marginal utility must rise forever and the resource stock will only be exhausted in the limit. If we instead assume that marginal utility is bounded at zero ($c < U'(0) < \infty$), then the resource stock will be exhausted in finite time at the precise instant that the rising marginal utility path reaches its upper bound. This is not only the planner's optimal extraction path, but it is also the aggregate extraction path that emerges in the competitive equilibrium of a decentralized market.

The main text discusses two reasons why a planner with the extraction technology described in section 3 would be unlikely to generate Hotelling's extraction path. First, oil flow in our model is constrained such that, even if the planner drilled every well immediately and produced at the maximum possible rate, oil would flow forever. Thus, price cannot rise at the rate of interest whenever oil is flowing, as Hotelling's path requires, unless marginal utility is also able to rise forever. Second, in our model, the incentive to drill in a given period depends, in part, on the cost of drilling in *other* periods, as captured by the $d'(a(t))\dot{a}(t)$ term in equation (17). One implication is that oil flow in our model can *increase* over intervals during which the marginal cost of drilling is falling, whereas the cost of extraction typically increases with production in standard models.

To overcome the first of these threats, we must assume that marginal utility is unbounded: $U'(0) = \infty$. To overcome the second, we must assume that the marginal cost of drilling is constant: $d(a) = \bar{d}$ for all $a \geq 0$. In this case, the imputed per-barrel cost of drilling is well-defined and is given by $c = \bar{d}(r + \lambda)/X$, which we assume is equivalent to the per-barrel extraction cost faced by Hotelling's planner. Given these two assumptions, condition (17) implies that the marginal utility of oil flow minus the per-barrel marginal cost of extraction rises at the rate of interest:

$$U'(F(t)) - c = \frac{\lambda \gamma_0}{X} e^{rt}. \tag{39}$$

Note, however, that we have implicitly assumed that the planner starts drilling wells at the outset of the planning period and never stops, producing at the constraint throughout, so that condition (17) always applies. For this to be the case, however, we need two other conditions to hold. First, the flow from drilled wells must decay sufficiently fast, so that the planner is able to achieve the Hotelling path via judicious control of the drilling rate while producing at her constraint. Second, the initial flow capacity K_0 cannot be too high, for otherwise the planner would delay drilling the first well—and may even produce below her constraint initially. If either of these conditions fails, then the planner is geologically constrained to an inferior path.

To illustrate the first of these two conditions, we assume for simplicity that drilling costs are zero ($c = 0$) and that utility from oil flow takes the constant elasticity form: $U(F) = \alpha F^{\beta}$ for $\alpha > 0, \beta \in (0, 1)$.[85] We also assume provisionally that $\lambda(1 - \beta) > r$. The total resource stock is given by $Q_0 = (K_0 + R_0 X)/\lambda$. In this case, the optimal program is given by:

$$\theta(t) = \frac{\gamma_0 e^{rt}}{X}, \text{ where } \gamma_0 = \frac{\alpha\beta X}{\lambda}\left(\frac{rQ_0}{1-\beta}\right)^{\beta-1} \tag{40}$$

$$F(t) = \frac{rQ_0}{1-\beta} e^{-\frac{r}{1-\beta}t} \tag{41}$$

$$a(t) = \frac{\left(\lambda - \frac{r}{1-\beta}\right)}{X} F(t) \tag{42}$$

$$\gamma(t) = \gamma_0 e^{rt}. \tag{43}$$

Note that the planning period begins with a pulse of drilling such that flow capacity totaling $rQ_0/(1 - \beta) - K_0$ is immediately added to the inherited capacity. Equation (42) implies that $a(t) > 0$ for all $t \geq 0$, since $F(t) > 0$ and since we have provisionally assumed that $\lambda(1 - \beta) - r > 0$.[86] It is straightforward to verify that this program satisfies each of the necessary conditions (7)–(13) and is therefore optimal for our planner.[87] Since $a(t) > 0$, it achieves the same discounted utility as Hotelling's planner.

However, our planner cannot always accomplish this feat. Suppose that the rate of decay from drilled wells is too low, with $\lambda(1-\beta) < r$. To achieve the result in (17) that $U'(F(t)) = \frac{\lambda\gamma_0}{X} e^{rt}$, the planner must set $\dot{F}(t) = \frac{rU'(F(t))}{U''(F(t))} = -\frac{rF(t)}{1-\beta}$. But then equation (12) implies that $a(t)X = \lambda F(t) - \frac{rF(t)}{1-\beta}$. Substituting and simplifying we conclude that $a(t)X = \frac{F(t)(\lambda(1-\beta)-r)}{1-\beta}$, which violates nonnegativity of $a(t)$. Intuitively, the planner is geologically constrained to a price path that rises more slowly than the rate of interest. Even if she drills all of the wells immediately and produces at the maximum possible rate, oil cannot be extracted quickly

[85] Note that η, the inverse oil demand elasticity (which we treat as a positive number), is given by $1 - \beta$. Also note that the β used in this section is a utility parameter, not the CAPM β used elsewhere in the main text and appendices.

[86] Since we have assumed that drilling and extraction are both costless in this example, any alternative drilling path such that the production path in (41) is feasible is also optimal. For any positive drilling cost, however, the planner would produce at the constraint and defer drilling until necessary.

[87] Equations (40) and (43) imply that condition (9) holds with equality. Equations (40) and (41) imply that condition (7) holds. Equation (43) ensures that (11) holds. Finally, equations (41) and (42) ensure that (12) holds.

enough to satisfy Hotelling's rule.

So suppose instead that $\lambda(1-\beta) > r$ and that the inherited flow capacity is too high, with $K_0 > rQ_0/(1-\beta)$. In this case, the planner will choose not to drill any wells initially, and production will decline at rate λ until drilling commences (when $K(t) = rQ(t)/(1-\beta)$). During this time, price will rise at a rate greater than r, and our planner will therefore not achieve the same utility as Hotelling's planner.[88]

To summarize, when the extraction technology involves drilling wells rather than producing barrels, we should not expect the Hotelling path to be optimal, unless four conditions hold: (1) the marginal cost of drilling a well is constant, (2) marginal utility is unbounded at zero, (3) the decay in flow from drilled wells is sufficiently fast, and (4) the inherited rate of oil flow is not too high. While the last two conditions seem reasonable (given that new wells are constantly being drilled in the real world) the first two conditions are not tenable. Our analysis of the Texas data shows clearly that marginal costs rise with the rate of drilling, while the viability of alternative fuels at *current* oil prices argues against an unbounded oil price.

[88]Since the planner can produce below the production constraint, the standard Hotelling path can still be achieved. However, this possibility requires either that drilling is costless or that it is so expensive that no wells are ever drilled. Otherwise, it can be shown that, given $d'(a) = 0$, price would initially rise at the rate of interest (while production is below the constraint and drilling is zero), would then rise faster than the rate of interest (after the constraint starts to bind while drilling remains at zero), and finally would rise more slowly than the rate of interest (after drilling turns positive with the constraint continuing to bind).

E Extending the drilling model to allow for costly above-ground storage

We have shown that our dynamics imply that it is possible, on the equilibrium path, for price temporarily to rise at a rate faster than the rate of interest while production is constrained. However, this result has not taken into account the possibility that oil may be stored above-ground. This appendix therefore asks how the availability of costly above-ground storage would affect our conclusions.[89]

For tractability, we assume that storage costs take the form of an iceberg storage cost—that is, a continuous, proportional erosion at rate m of the quantity (and therefore the value) of stored oil. In this case, a stockpiler receiving capital gains but no convenience yield would be indifferent between buying or selling oil if the oil price rises at the constant percentage rate of $m + r$.

Importantly, this extended model still accommodates the possibility that the oil price can rise more quickly than r on the equilibrium path while production is constrained. In fact, during any measurable interval over which storage is occurring (and therefore the oil price is rising at the rate $m+r$), production *must* be constrained. Why? As we have seen, whenever production is unconstrained, the oil price must be rising at r, not $r + m$. Hence, even if above-ground storage is possible, the equilibrium may still involve intervals over which price rises in percentage terms at rate $r+m$, and production must be at capacity throughout each of these intervals. Such a steep rise in prices was actually anticipated in winter 1998–1999 (see figure 3) and was in fact accompanied by a surge in above-ground storage (see figure 14 in appendix A).

However, as figure 14 indicates, above-ground storage also occurred when expected capital gains were significantly smaller. This suggests that above-ground storage also provides a convenience yield that makes carrying inventory attractive to at least some stockpilers even when capital gains do not cover their interest and storage costs.[90] If we amend our account to include heterogeneous stockpilers, each with a convenience yield that is strictly concave in the amount stored, then above-ground storage would still put a ceiling on capital gains of $r + m$ and would still be accompanied by production at the constraint whenever the percentage change in the oil price differs from r. However, because of the convenience yields, storage could occur even when holding inventory results in net capital losses.

E.1 Necessary conditions and their implications

We now formally present the Hamiltonian-Lagrangean corresponding to the social planner's problem with costly storage and derive and interpret the implied necessary conditions. With storage, let the choice variable $C(t)$ denote oil consumption generating instantaneous utility

[89] It is clear that, if storage were costless and limitless, then the oil price could never rise more quickly than r. Real-world logistical costs and storage constraints argue against this hypothetical, as does the fact that we observe periods in our data (using either unadjusted or risk-adjusted futures prices) when the price is expected to rise more quickly than r.

[90] A convenience yield naturally arises in our setting from the fact that many leases are not connected to the pipeline network but instead sell their produced oil via tanker truck. Firms must store produced oil in on-site storage tanks between truck pick-ups.

flow of $U(C(t))$. Let $S(t) \geq 0$ denote a new state variable representing the quantity of oil stored at time t. The equation of motion for S with an iceberg storage cost of m is then $\dot{S}(t) = F(t) - C(t) - mS(t)$, where $F(t)$ is still the choice variable representing oil production, as above. The other variables and constraints—the drilling rate and drilling costs, production capacity and its equation of motion, and the stock of remaining wells and its equation of motion—are all the same as above.

The current-value Hamiltonian-Lagrangean of this maximization problem is then given by:

$$H = U(C(t)) - D(a(t)) + \theta(t)[a(t)X - \lambda F(t)] + \gamma(t)[-a(t)] + \phi(t)[K(t) - F(t)] + \mu(t)[F(t) - C(t) - mS(t)], \quad (44)$$

where $\mu(t)$ is a new co-state variable on $S(t)$.

Necessary conditions are given by the following (renumbering some conditions that match those from the original problem without storage):

$$C(t) \geq 0,\ U'(C(t)) - \mu(t) \leq 0,\ c.s. \quad (45)$$
$$F(t) \geq 0,\ \mu(t) - \lambda\theta(t) - \phi(t) \leq 0,\ c.s. \quad (46)$$
$$F(t) \leq K(t),\ \phi(t) \geq 0,\ c.s. \quad (47)$$
$$a(t) \geq 0,\ \theta(t)X - d(a(t)) - \gamma(t) \leq 0, c.s. \quad (48)$$
$$\dot{R}(t) = -a(t),\ R_0 \text{ given} \quad (49)$$
$$\dot{\gamma}(t) = r\gamma(t) \quad (50)$$
$$\dot{K}(t) = a(t)X - \lambda F(t),\ K_0 \text{ given} \quad (51)$$
$$\dot{\theta}(t) = -\phi(t) + r\theta(t) \quad (52)$$
$$\dot{S}(t) = F(t) - C(t) - mS(t),\ S_0 = 0 \quad (53)$$
$$\dot{\mu}(t) \leq (r+m)\mu(t),\ S(t) \geq 0,\ c.s. \quad (54)$$
$$K(t)\theta(t)e^{-rt} \to 0,\ R(t)\gamma(t)e^{-rt} \to 0, \text{ and } S(t)\mu(t)e^{-rt} \to 0 \text{ as } t \to \infty. \quad (55)$$

These necessary conditions differ in two main ways from those presented above. First, we now have necessary conditions that characterize production and consumption incentives separately. Condition (45) says that the price of oil will equal the shadow value on storage capacity whenever consumption is strictly positive and will be strictly less than this value when consumption is zero. Meanwhile, condition (46) is the same as in the main text, but with the price of oil replaced by the shadow value of storage $\mu(t)$. When consumption is strictly positive, then $\mu(t)$ can be eliminated to yield the same condition characterizing production incentives that we had before. Second, the necessary conditions now include (53), which is the new equation of motion for storage, along with condition (54), which characterizes the evolution of its co-state variable. When consumption is strictly positive, this latter condition says that price must rise at rate $r + m$ during any interval over which storage occurs. Thus, during any such interval, the production constraint must bind.

In Anderson et al. (2014), we show that consumption is always non-zero in this model (assuming that drilling at a sufficiently low rate is profitable when starting from an initial

capacity $K_0 = 0$). We also show that once production is capacity-constrained, it must always subsequently be constrained, and if initial capacity is sufficiently low, production is always constrained along the equilibrium path. Intuitively, storage offers a substitute means by which consumption can be deferred to the future, so introducing storage into the model will not increase the incentive to produce below the available capacity constraint.

F Details for the computational model in section 6.1

Section 6.1 introduces computationally solved drilling, extraction, and price paths for a specific case of the model in equilibrium. This appendix discusses the details of the computation procedure.

We obtain an approximate computational solution to the model by optimizing the social planner's welfare function using value function iteration. The model has two state variables, R and K. Because the demand and drilling cost functions we specify satisfy the sufficient conditions discussed in Anderson et al. (2014), we have that $F = K$ so that there is only one choice variable, a.

A fairly dense state space is required to yield smooth-looking simulations of drilling, extraction, and prices. The state space for R consists of 600 steps, while that for F consists of 300 steps (where the maximum value for F is given by $X * R_{max}/4$). The action space for a consists of 600 steps (where the maximum value for a is given by $R_{max}/40$). Because numerical precision is particularly important when a, R, and F are small, we compress the discrete actions and states at the lower end of their support by making the step lengths equal in the square root of each state variable. For both solving and simulating the model, we use a time step of three months.

In the value iteration loop, for each iteration we calculate for each discrete state s the optimal discrete action a. Because each possible action is very unlikely to cause a discrete state to be hit exactly in the next period, we obtain the value for next period's state via linear interpolation. For value function convergence, we use a tolerance of 10^{-7} on the sup norm (with the value function measured in units of $million).

The solution to the model yields a policy function for the rate of drilling at all possible discrete states. To conduct the forward simulation, we linearly interpolate this policy function whenever the simulation requires the optimal drilling rate for a state that does not exactly match one of the discrete states.