

**Enhancement of unconventional oil and gas production
forecasting using mechanistic-statistical modeling**

by

Justin B. Montgomery

B.S. Mechanical Engineering, Texas A&M University (2013)

S.M., Technology and Policy, Massachusetts Institute of Technology (2015)

Submitted to the Department of Civil and Environmental Engineering and
the Center for Computational Engineering

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2020

© Massachusetts Institute of Technology 2020. All rights reserved.

Signature redacted

Author

Department of Civil and Environmental Engineering

Signature redacted

January 15, 2020

Certified by



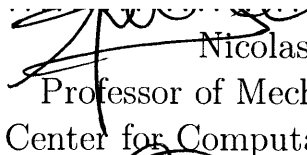
John R. Williams

Professor of Civil and Environmental Engineering

Signature redacted

Thesis Supervisor

Accepted by



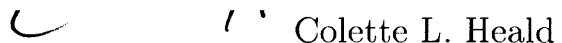
Nicolas Hadjiconstantinou

Professor of Mechanical Engineering

Director, Center for Computational Engineering

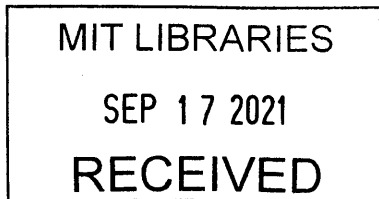
Accepted by

Signature redacted


Colette L. Heald

Professor of Civil and Environmental Engineering

Chair, Graduate Program Committee



ARCHIVES

Enhancement of unconventional oil and gas production forecasting using mechanistic-statistical modeling

by

Justin B. Montgomery

Submitted to the Department of Civil and Environmental Engineering on January 15, 2020,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

Unconventional oil and gas basins have rapidly become expansive and critical energy resource systems. However, accurately predicting highly variable well production rates remains challenging, given the typically poor subsurface characterization and complex flow behavior involved. This creates uncertainty about future resource availability, undermining reliable economic assessments and good stewardship of the resource.

Production, drilling, and hydraulic fracturing datasets from thousands of wells offer insight into patterns of productivity but are noisy and incomplete. Fully exploiting this information is only possible by leveraging contextual knowledge to structure observations. This thesis provides a novel framework for combining machine learning and probabilistic modeling with domain knowledge and physics to understand and predict well productivity.

Technology is a constantly evolving driver of productivity that must be captured in forecasts. This thesis shows that the immense geological heterogeneity of unconventional basins can lead to overestimating the role of technology when the best areas are increasingly targeted alongside design improvements. This conflation is remedied using spatial structure to infer geological productivity as a latent variable. A regression-kriging technique is shown to effectively disentangle technology from geology—which play roughly equal roles—and reduce error in initial well productivity predictions by more than a third compared to established methods.

Long-term production dynamics for unconventional wells are unpredictable and current forecasting approaches have considerable limitations. Fitted production curve models are ill-posed and unreliable, but aggregated type-well curves ignore important differences between wells. This thesis introduces Tikhonov regularization as a way of effectively sharing information across wells, cutting error in the earliest long-term productivity forecasts in half. Additionally, a spatiotemporal hierarchical Bayesian approach is developed that incorporates physical relationships to enhance predictions and interpretability while quantifying and reducing uncertainty. Sampling from this high dimensional model is enabled by designing a unique Metropolis-Hastings within Gibbs scheme to take advantage of the model's structure. This novel mechanistic-statistical approach is able to learn and generalize physical relationships across ensembles of wells with vastly different properties—realistic scenarios where current techniques generate two to five times as much error—providing an important and practical advance in better understanding and managing these resources.

Thesis Supervisor: John R. Williams

Title: Professor of Civil and Environmental Engineering

Acknowledgments

I want to acknowledge and thank the many people who have contributed to this research and supported me throughout my PhD. I am fortunate to have had such a fantastic doctoral committee, including my faculty advisor and coauthor for Chap. 3, John Williams; research funding supervisor and coauthor for Chap. 2 and Chap. 3 (as well as master’s thesis supervisor, mentor, and friend), Francis O’Sullivan; committee chair, Ruben Juanes; and coauthor for Chap. 4 and guide for all things Bayesian, Youssef Marzouk. They have consistently challenged me while providing me with the thoughtful, multi-disciplinary guidance that I needed.

Beyond my committee, many friends and members of the MIT community have contributed their time and expertise to assist me in my research. Andrew Davis was a coauthor for the research in Chap. 4 and always found the time to help me navigate the wild world of Markov chain Monte Carlo. Samuel Raymond was a coauthor for Chap. 3 and consistently a great sounding board for ideas about machine learning in physical systems. They both helped me to expand the breadth of this research without losing any depth. Additionally, Emre Gençer was an ongoing source of guidance and support in the later stages of my research. Others at MIT shaped my research through wonderful discussions and advice, including Aimè Fournier, Gordon Kaufman, Simone Cenci, and Seonkyoo Yoon.

I am grateful for the important industry feedback and suggestions I received from Hunter Hunt, Victor Liu, and others at Hunt Energy; Robert Kleinberg (Schlumberger and Presidio Energy Technology), John Tolle (Shell), Dan Shaughnessy (interp3), Behrouz Ebrahimi (Halliburton), and too many others to mention here. Ongoing industry input was critical to making this research practical. Additionally, the engagement I had with members from the U.S. Energy Information Administration including Faouzi Aloulou and Troy Cook was tremendously valuable to this research. I also want to thank several academics outside of MIT: Rob Perrons of Queensland University of Technology for his mentorship; Frederik vom Scheidt and others at Karlsruhe Institute of Technology for graciously hosting me as a visiting scholar; Svetlana Ikonnikova, Scott Hamlin, and Guinevere McDaid of the Texas Bureau of Economic Geology for providing shapefiles of their published geological maps used

in Chap 4.

I am grateful to the MIT Energy Initiative (MITEI) for funding and hosting me as a PhD Researcher. It has been a pleasure to share an office with the many wonderful people that support and carry out research in MITEI; I especially want to thank Lou Carranza for always promoting my research and Jenn Schlick for her help launching shalestats, an online interactive economic modeling tool for North Dakota wells based on the results of Chap. 3. Exxon-Mobil supported my final months of research and I thank them for this.

My time at MIT has been both challenging and also incredibly enjoyable thanks to the many wonderful people I have shared this time with. I am thankful to the Civil and Environmental Engineering department for backing the crazy idea of a department rock band and for the bandmates that helped me to make this a raging success.

Finally, I want to thank the most important people in my life: my parents, my siblings Nicole and Cody, and my uncle Jeff for always supporting and believing in me; my lifelong friends Alex, Jerod, and Jackson for getting me through the emotional ups and downs of this time; and last (but definitely not least!), Myriem for simply being amazing and giving my PhD a happy ending.

Contents

1	Introduction	15
1.1	Background and motivation	15
1.2	Unconventional production forecasting: challenges and opportunities	16
1.3	Central themes	18
1.4	Contribution and overview	20
2	Disentangling the role of drilling location from evolving technology	23
2.1	Introduction	24
2.2	Data and Methodology	28
2.2.1	Data	28
2.2.2	Multiple linear regression	29
2.2.3	Omitted variable bias and spatial autocorrelation	31
2.2.4	Regression models accounting for spatial autocorrelation	34
2.2.5	Evaluating the results	36
2.3	Results and discussion	38
2.3.1	Regression model estimates	38
2.3.2	Forecasting applications	41
2.3.3	Dis-aggregating the productivity trend	45
2.4	Concluding remarks	48
3	The fundamental ambiguity in mechanistic-statistical production forecasting	51
3.1	Introduction	52

3.2	Analysis	55
3.3	Discussion	59
4	A hierarchical Bayesian approach to incorporate physical information into mechanistic-statistical production forecasting	65
4.1	Introduction	65
4.2	Methods	72
4.2.1	Least-squares and type-well curves	72
4.2.2	Bayesian Formulation for Production Curve Models	72
4.2.3	Hierarchical formulation with Gaussian process	74
4.2.4	Sampling using Metropolis within Gibbs	77
4.3	Application	83
4.3.1	Data	85
4.3.2	Results	87
4.4	Conclusions	98
5	Conclusion	103
5.1	Summary	103
5.2	Future work	106

List of Figures

2-1	Productivity of drilling rigs and new wells in Williston basin	25
2-2	Predicted and actual values for four of the regression models	40
2-3	Comparison of average spatial weights used by spatial models	41
2-4	Regression parameter estimates and confidence interval	42
2-5	The mean and interquartile range of actual well productivity over time and the mean prediction for each model	43
2-6	The mean of actual well productivity over time and the mean prediction for each model based only on early data	43
2-7	Forecast of mean first year production for wells drilled in the first half of 2015 with 2018 design parameters	44
2-8	Predictions of first year production for locations within the Bakken formation in North Dakota using design parameters expected for 2018	46
2-9	Comparison of predicted mean well productivity to predictions with technol- ogy held constant for different regression models	47
2-10	Breakdown of the relative influence of factors on the productivity improvement	48
3-1	A comparison of prediction accuracy for 10-year cumulative production with different approaches and months of production data.	56
3-2	Sloppiness: Ratio of the largest to smallest eigenvalues of the Hessian	58
3-3	The loss function associated with fitting the SCM to well production	61
3-4	The loss function associated with fitting the modified Arps curve to well pro- duction plotted against b and D_i	62

3-5	The loss function associated with fitting the modified Arps curve to well production plotted against b and Q_i	63
4-1	Horizontal well and planar fracture geometry assumed in the simplified one-dimensional flow model	68
4-2	LGM fit using nonlinear least squares to realization of SCM with typical parameters	70
4-3	Parameter mappings between the SCM and LGM	71
4-4	MCMC samples of coefficients for regression coefficients and first three modes of KL expansion with Barnett wells	87
4-5	Comparison of coefficients using different models with Bakken wells	88
4-6	Map of wells and expected long-term well productivity in Barnett field with Lateral length held to average value	89
4-7	Map of wells and expected long-term well productivity in Bakken field with design parameters held to average values	90
4-8	Hierarchical model fit to observed production in first year for a Barnett well giving posterior predictive mean and credible intervals for later production	91
4-9	Hierarchical model fit to observed production in first year for a Bakken well giving posterior predictive mean and credible intervals for later production	92
4-10	Two-fold cross validation results of hierarchical model posterior predictive and actual ten-year cumulative production for Barnett wells	93
4-11	Two-fold cross validation results of hierarchical model posterior predictive and actual five-year cumulative production for Bakken wells	93
4-12	Root mean squared error and average error for two-fold cross validation predictions of Barnett wells' ten-year cumulative production	94
4-13	Root mean squared error and average error for two-fold cross validation predictions of Bakken wells' five-year cumulative production	94
4-14	Uncertainty for Barnett wells' ten-year cumulative production in two-fold cross validation using hierarchical model and type-well curve	95

4-15 Uncertainty for Bakken wells’ five-year cumulative production in two-fold cross validation using hierarchical model and type-well curve 95

4-16 Training and test wells in Bakken shale chosen based on hydraulic fracturing water volume 96

4-17 Root mean squared error and average error of predicted five-year cumulative production for Bakken wells with largest stimulations 96

4-18 Training and test wells in Barnett shale chosen based on lateral length and reservoir quality 97

4-19 Root mean squared error and average error of predicted ten-year cumulative production for Barnett wells with longest laterals and in highest-quality reservoir 98

4-20 Relative change in error by using empirical Bayesian approximation instead of mechanistic-statistical hierarchical model 99

List of Tables

2.1	Regression models used	37
2.2	Comparison of performance for the regression models	39
2.3	Distribution of predicted productivity for wells drilled in the first half of 2015 with 2018 design parameters	45
4.1	Data and models for Barnett and Bakken	86

Chapter 1

Introduction

1.1 Background and motivation

The rapid growth of production from unconventional shale and tight resource systems has turned the United States into the world's leading producer of both natural gas and crude oil. These resources were long neglected due to the ultra-low permeability of the rocks. However, attractive production rates were enabled by designing wells that maximize contact with the reservoir through long horizontal sections combined with massive water-based hydraulic fracturing stimulations containing proppant (usually sand particles) to ensure fractures remain propped open.

Widespread development activity in these fields has been accompanied by concerns for the enormous environmental footprint of wells, including the acquisition and disposal of immense volumes of water for hydraulic fracturing and the emissions associated with production [53, 13]. Abundant and cheap natural gas from shale resources has drastically altered the economics of electricity generation in the United States, leading to a shift away from coal and reduction in carbon dioxide emissions [53]. Nevertheless, substantial debate continues over construction of new pipelines to transport the burgeoning resource due to the enduring dependence this locks in for carbon emitting fuels and persistent skepticism about the long-term economic viability [71].

Already, the economics of the resource have begun to face greater scrutiny as investors have grown increasingly wary of the capital intensive nature of development combined with

elusive returns due to unpredictable and highly variable production from wells [77]¹. Although overall production has proved exceptionally resilient and even increased following price collapses in natural gas and then oil, many wells have turned out to be unprofitable, under-performing operators' production projections [3, 83]. Behind this sits a fundamental challenge with the resource: production rates from unconventional oil and gas wells are difficult to predict, even with wells that appear to be quite similar, and the physical drivers of well productivity are still poorly understood.

Improving upon the techniques used today to develop forecasts and understand the mechanisms of production will unlock tremendous value by more effectively informing decisions related to the resource's development. The implications of this for the public range from advancing our understanding of the environmental impact of development to informing energy policy decisions and massive infrastructure investments. In the private sector, the impact will be more efficient deployment of capital—the dominant concern for operating companies today—with a more realistic understanding of production levers enabling further optimization of development campaigns and well designs. Additionally, more reliable production forecasts will allow for better valuation and risk assessment of investments in this area.

1.2 Unconventional production forecasting: challenges and opportunities

Traditional physics-based reservoir simulations are of limited utility with unconventional oil and gas wells due to the cost and complexity of characterizing and modeling the complex subsurface conditions and nanoscale flow behavior [61, 68]. Pore diameters in shale rocks are of a similar magnitude to the mean free path of molecules, making molecule-wall interactions non-negligible. It is difficult to characterize the heterogeneous properties of the reservoir and capture nonlinear behavior due to pressure changes, including desorption and pore size changes [110]. Additionally, there is immense uncertainty about the placement and properties

¹An extensive discussion of the nature of uncertainty and risk in unconventional oil and gas can be found in my master's thesis: *Characterizing Shale Gas and Tight Oil Drilling and Production Performance Variability* [77]

of the fractures that are key to productivity [49, 109].

As a result, forecasting by both industry and government today is dominated by data-driven techniques [105, 93]. However, many of these approaches are unable to adequately address the geological heterogeneity of unconventional reservoirs and the constantly evolving technology involved. This nonstationarity makes it difficult to use past performance of older offset wells as production analogues for new wells developed under vastly different conditions. Many of the approaches used today also fail to rigorously capture the uncertainty inherent in available data and convey this in the forecasts being generated. Furthermore, these approaches are unable to effectively incorporate the wealth of domain knowledge, including geological and petroleum engineering principles involved in production, and the abundance of field data from other producing wells.

Data from these unconventional fields is indeed prolific but of a fundamentally different kind and quality than from the conventional fields that have long been targeted, making new modeling tools a necessity. Large conventional drilling prospects have typically been extensively logged and studied prior to development since costs can be immense and the primary risk is of not finding a petroleum deposit at all [77]. Additionally, because the physics of conventional production are well understood, there is a strong incentive to acquire this expensive subsurface information, which can be used in simulations to guide decisions over the potentially long lifetime of a well. By contrast, the more uncertain physics for unconventional wells—combined with their relatively quick payback time—makes it more attractive to put capital into drilling additional wells rather than trying to characterize the heterogeneous subsurface. Data available for an individual unconventional well is inadequate on its own to reasonably predict future behavior. It is thus necessary to glean as much as possible from the extremely noisy and incomplete data of the broader field, which is often available due to regulator reporting requirements, while enriching this understanding with more thorough measurements for subsets of wells.

In unconventional oil and gas basins, patterns of productivity are best understood by viewing the thousands of producing wells as a collection of ongoing small experiments. Instead of limiting an analysis to one well or a subset of similar wells, it is important to systematically share information across all wells to reduce uncertainty and enhance the over-

all understanding of production dynamics. Given the lack of measurements for some critical properties of wells, it is also essential to structure data based on prior knowledge in order to better separate signal from noise and infer latent subsurface conditions from the entire population. There is considerable interest today in applying machine learning techniques to data from unconventional fields. The goal of this thesis is a more nuanced approach to this that strives to also retain interpretability and combine these powerful predictive tools with rigorous uncertainty quantification, physics and domain knowledge wherever possible.

1.3 Central themes

Within the methodological framework that this thesis provides, there are five central themes, or guiding philosophies, that should be recognized as critical considerations for modeling unconventional well productivity more generally. These address fundamental data and modeling issues within this context and are worth highlighting since they make the advantages of the approach more clear and suggest how insights from this thesis may be adapted to address slightly different questions about unconventional well productivity. Although the framework presented in the following chapters is flexible and powerful, the building blocks identified here suggest how it can best be adapted to the inevitable different datasets and decisions that arise. Furthermore, these themes help to connect the specific research application addressed here to broader principles for data-driven modeling of uncertain complex physical systems.

The themes are briefly introduced here and will be mentioned again in Chap. 5 to point out how these interrelated concepts permeate and tie together the different parts of this thesis.

- **Linear regression is simple but effective:** Using least squares fits of a linear model to describe relationships between variables is one of the simplest and most fundamental techniques for data analysis. Given the noisiness in unconventional oil and gas data and the need for interpretability to inform operational decisions, this technique should not be overlooked or underestimated. More complicated machine learning approaches may be seductive but it is always important to consider whether their complexity is

warranted for the quality of data available since they can easily overfit noisy data, leading to poor generalization with new data. The simplicity of linear regression also makes it easy to combine it with more sophisticated modeling techniques to address other important aspects of the system.

- **Latent variables should not be ignored:** Important physical attributes, especially subsurface properties, are usually unavailable for unconventional wells. This makes it important to draw on assumed structure for the system, known *a priori* or evident across large populations of wells, in order to infer them as latent variables. This can help to resolve some of the variability in production data, making relationships clearer and forecasts more accurate.
- **Balance the individual and the population using partial-pooling:** A common dilemma with current forecasting techniques is how to group wells for comparison and analysis. Should wells be treated individually, given their geological and technological heterogeneity, or as similar enough to be lumped together for analysis? In reality, a middle ground is often sought by modelers in which data is fully-pooled across a small subsample of neighboring wells deemed to be sufficiently homogeneous. However, this introduces an inherent tradeoff: Include only the closest wells and the small sample size is overwhelmed by the noisiness of the data; group wells too broadly and there is a risk of missing important differences that can contaminate the analysis. A common theme throughout this thesis is how partial-pooling of data using spatial structure, regularization, and hierarchical priors circumvents the drawbacks of all-or-nothing, completely pooled or unpooled approaches.
- **Ill-posedness requires regularization:** The most readily available data for producing wells is the observed production rates and it is common to attempt to infer subsurface conditions from this production data. However, this data on its own is inadequate since it can be explained by different combinations of subsurface parameters. Regularization augments a well's production data with some kind of additional information, such as production data from other wells in the field, and is essential for resolving this ambiguity.

- **Uncertainty quantification is essential:** In this context, where data is noisy and relationships can be modeled only very approximately, it is important to characterize the distribution of uncertainty in predictions. The large variance in unconventional well productivity means that deterministic estimates can be misleading and a poor basis for decision-making. Additionally, uncertainty may be asymmetric, as shown in [77], and risk can be distorted in unpredictable ways by nonlinear system behavior and aggregation over many wells. Rigorous uncertainty quantification requires treating quantities as probability distributions which are updated by new information, rather than as fixed values with some arbitrary range of confidence applied.

1.4 Contribution and overview

The main contribution of this thesis is a novel, data-driven methodology for better quantifying uncertainty and improving the accuracy in production forecasts while also enhancing understanding about the factors behind productivity. This is accomplished by developing modeling tools and a framework that allow existing disparate and imperfect sources of data and domain knowledge, including mechanistic models, to be systematically incorporated into forecasts. This is a significant advance to the area of unconventional production forecasting since the crude, heuristic methods currently relied upon are unable to leverage this abundant information and are less useful and accurate as a result.

Rather than an outright rejection of the forecasting methods currently used, this thesis will expose the flawed statistical assumptions of these approaches and demonstrate the benefits of introducing more sophisticated modeling techniques. The simpler methods currently used are indeed a form of domain knowledge and by improving and building on them, this thesis provides a methodology that is more practical and likely to be adopted by other analysts and forecasters. It also makes clearer the limitations and issues of continuing to rely on current methodologies. In fact, the ideas in Chap. 2 have already had the desired impact. Publication of these findings in [74] led to widespread acknowledgment of flaws in current modeling approaches at the U.S. Energy Information Administration [86, 75] and adoption of the new recommended technique by industry analysts at, for example, Rystad

Energy [88].

This contribution is developed in three stages (and chapters) and organized around the themes discussed in Sect. 1.3. First, in Chap. 2 a highly resolved spatial model for more accurately predicting initial production rates is developed with a central focus on how to disentangle the role of technology from that of geology. Next, in Chap. 3 the fundamental ambiguity associated with forecasting well production over long periods of time is revealed and regularization is shown to reduce ill-posedness and substantially boost accuracy. Finally, these ideas are combined in Chap. 4 using a hierarchical Bayesian model that quantifies uncertainty while providing the essential dose of regularization and capturing the influence of technology and geology. This basin-level model is able to learn subtle statistical and physical relationships across thousands of wells and embed this information into a mechanistic production model that can be used to help manage the increasingly complex challenges of unconventional development.

Chapter 2

Disentangling the role of drilling location from evolving technology¹

New well productivity levels have increased steadily across the major shale gas and tight oil basins of North America since large-scale development began just over a decade ago. These gains have come about through a combination of improved well and hydraulic fracturing design, and a greater concentration of drilling activity in higher quality acreage, the so called “sweets spots.” Accurate assessment of the future potential of shale and tight resources depends on properly disentangling the influence of technology from that of well location and the associated geology, but this remains a challenge. This chapter describes how regression analysis of the impact of design choices on well productivity can yield highly erroneous estimates if spatial dependence is not controlled for at a sufficiently high resolution. Two regression approaches, the spatial error model and regression-kriging, are advanced as appropriate methods and compared to simpler but widely used regression models with limited spatial fidelity. A case study in which these methods are applied to a large contemporary well dataset from the Williston Basin in North Dakota reveals that only about half of the improvement in well productivity is associated with technology changes, but the simpler regression models substantially overestimate the impact of technology by attributing location-driven improvement to design changes. Because of the widespread reliance on these less spatially

¹This chapter is based on the article *Spatial variability of tight oil well productivity and the impact of technology* [74]

resolved regression models, including by the U.S. Energy Information Administration to project shale gas and tight oil resource potential, the overestimate of technology’s role in well productivity has important implications for future resource availability and economics, and the development choices of individual operators.

2.1 Introduction

Oil and gas produced from shale and tight rock formations is playing an increasingly important role in global and domestic energy markets. Due to increased production of oil from North Dakota, Texas, and other states, the United States is now considered by some to be the world’s “swing producer,” supplanting OPEC in this traditional oil market balancing role [1]. In North Dakota, which includes the most active part of the Williston tight oil basin, crude oil production grew from 98 thousand barrels a day (Mbbbl/d) in 2005 to 1174 Mbbbl/d a decade later. Additionally, the U.S. power sector has drastically increased its reliance on domestically produced natural gas, especially from shale [101]. Although these formations have long been known to contain abundant oil and gas, the “tightness,” or low permeability, of the rock led many to view production from them as not economically viable [54]. However, commercial rates of production turned out to be possible using long horizontally-drilled wells combined with hydraulic fracturing—in which fluid and sand is pumped into wells to break apart rock and create pathways for fluid flow—and this has led to the rapid expansion of shale gas and tight oil production in the past decade [79, 52, 56].

In recent years there has been a sustained downturn in oil and gas prices, leading to substantial uncertainty about future levels of production from shale gas and tight oil formations [2]. The future outlook for these resources now depends largely on the capacity of industry to improve the economics of extraction through higher productivity. Thus far there have been signs of this happening, with production per drilling rig increasing as the number of active drilling rigs has fallen precipitously, as shown for the Williston basin in Fig. 2-1(a) [102]. Although some of this trend can be attributed to more efficient drilling, much of it is driven by a rise in average new well productivity (Fig. 2-1(b)) [77, 56, 78].

There are two important factors to recognize behind increases in well productivity. First,

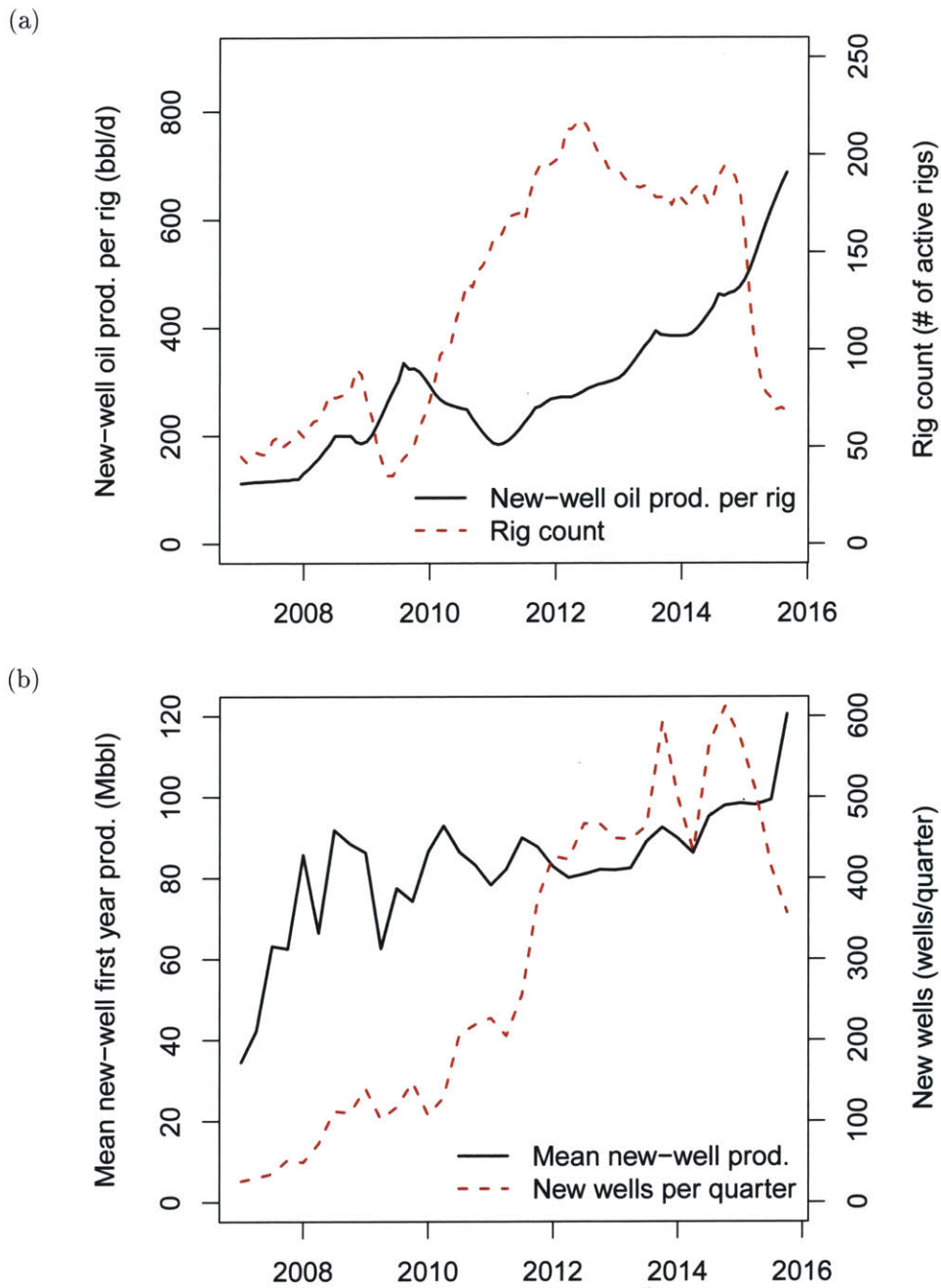


Figure 2-1: Two perspectives on productivity in Williston tight oil basin. (a) Productivity of drilling rigs, measured as production per active rig (Source: U. S. Energy Information Administration [102]). (b) Productivity of new wells, measured as mean first year production of new wells in each fiscal quarter.

oil and gas operating companies have been scaling up well designs in an effort to increase well production through greater reservoir access. There has been a shift toward longer lateral lengths—which tends to mean a greater number of hydraulic fracturing “stages” at which fractures are initiated from—and larger volumes of both the water-based fracturing fluid pumped to create fractures and the sand, or proppant, carried by this fluid in order to keep the fractures “propped” open after water has flowed back [108, 56, 7, 100, 34]. In addition to this, operating companies have been “high-grading,” or focusing their drilling efforts on the locations in a field with the most favorable geology and highest expected production [46, 57, 21, 89, 100]. Technology-driven improvements to productivity may be transferable to future wells in all parts of a field but high-grading amounts to simply exploiting the lowest-cost resource first. To understand changes in resource economics and realistically forecast future production it is therefore critical to be able to distinguish accurately between the influence of location and technology choices on well productivity [24, 44].

Multivariate statistical analysis remains an important approach to understanding the role that technology choices have played on well productivity. There are large datasets of production and engineering data available, due to the large number of wells that have already been drilled in these formations [84, 22, 15]. Additionally, there are limitations to physics-based modeling approaches due to frequently inadequate well-level geological data and the challenges of simulating fracture propagation and complex flow behavior in low permeability rock [69, 20]. As a result, multivariate regression modeling has been widely adopted to infer the impact of technology on tight oil and shale gas well productivity [40, 60, 35, 66, 17, 115, 51, 101, 27, 22, 106].

An important modeling challenge associated with this is how to control for location, since reservoir quality, and hence well productivity, is spatially dependent. Some authors have chosen to simply ignore this feature and use *nonspatial* models, but this makes it unclear how reliable their results are [40, 60, 35, 66]. At the other end of the spectrum, location or functions of location can be included as independent variables in a regression model, using *surface trend analysis* [27]. Another approach to control for location lies in between these, and assumes geological homogeneity within a small sample of wells [17, 115, 51], or within *fixed effects* regions [101, 27, 22, 106]. For example, the U.S. Energy Information

Administration (EIA) assumes county-level fixed effects, in which the difference in each well's productivity from the mean in its county is attributed to the influence of technology.

Implicit in all of these approaches is an assumption that spatial variability can be neglected below some arbitrary scale and this assumption will not overly influence results. However, important properties in shale and tight reservoirs have been found to vary considerably over even relatively short distances [20]. Furthermore, the tendency of operating companies to high-grade drilling activity alongside the scaling up of technology parameters creates a risk of conflating these impacts and potentially under- or over-estimating the amount of technological improvement actually made. No study has specifically considered the potential of different controls for location to influence inference results and it is difficult to compare estimates between studies since different datasets and assumptions have been used. Studies that have adopted some controls for location have generally concluded that differences in well location play an important role on well productivity, but a lack of robust controls for location has made it difficult to quantify this relationship in the past [51, 27, 112, 47, 106].

In other domains with spatially dependent data, such as ecology, soil science, and urban energy consumption, regression-kriging and spatial error models have been used to explicitly incorporate spatial autocorrelation, or the spatial clustering of similar observations, into estimates [26, 12, 39, 97]. These approaches have not yet been previously used to distinguish between the influence of location and technology on tight oil and shale gas well productivity. This chapter will apply these approaches to a large dataset of wells from the Williston tight oil basin in order to rigorously quantify the impact of changes in technology and location over a 42 month period. Additionally, three models which appear to be the current standard—a nonspatial model, fixed effects model, and surface trend analysis model—will be compared in order to understand how influential the choice of spatial controls in a well productivity regression model is for estimates. Section 2.2 discusses the data and statistical methods used in the analysis. The models which are currently use in this area are discussed in Section 2.2.2 and their potential for biased estimates in Section 2.2.3. The models accounting for spatial autocorrelation are described in Section 2.2.4. In Section 2.3, the results of the five different models are compared and discussed, including a comparison of near-term forecasts in Section 2.3.2 and finally a breakdown of the relative contribution of technology and location

in Section 2.3.3. Finally, important conclusions are highlighted in Section 2.4.

2.2 Data and Methodology

2.2.1 Data

The data used in this study comes from horizontal wells drilled into either the Middle Bakken or underlying Three Forks formations in North Dakota during a 42 month period beginning in 2012. These formations are the primary oil and gas producing layers in the Williston basin, a large sedimentary depression spanning North Dakota, South Dakota, Montana, and Saskatchewan [85]. The most productive and actively developed region is near the center of the basin in north-western North Dakota [14]. This is the area used for this study because of the prolific drilling activity in recent years, and the quality and availability of public well data with uniform reporting standards. Additionally, there are strong spatial trends in productivity across this area, with a sweet-spot located generally in Mountrail and McKenzie counties and diminishing productivity moving outward.

A fully assembled dataset of North Dakota wells for this analysis was obtained from Drillinginfo, a data provider (web: drillinginfo.com). This well-level data was reported by operating companies to the North Dakota Department of Mineral Resources (web: dmr.nd.gov) and included monthly production rates, perforated lateral lengths, date of first production, and drilling location geographic coordinates. Data for the total mass of proppant and volume of water used for hydraulic fracturing was drawn from Frac Focus (web: fracfocus.org), a hydraulic fracturing chemical registry with mandatory reporting in 23 states, including North Dakota. Additionally, structural contour and isopach shapefile maps from the North Dakota Geological Survey were used to identify well target formation based on reported vertical depth measurement [62, 63]. By combining these data sources, a dataset was compiled for 3644 wells with complete reporting of well design parameters of interest and at least a year of production.

The well and hydraulic fracturing design, or technology parameters included in this analysis were well lateral length, total water, and total proppant. These have been previously

identified to be among the most influential design variables for well productivity [17, 51, 60, 65, 66, 73, 90, 91, 106, 115]. Some additional operational parameters that have been identified as relevant to well productivity include the number of hydraulic fracturing stages and the orientation of the well relative to *in situ* stress of the formation [52, 85, 89, 7, 91]. However, by the date of the earliest wells analyzed, it is likely that operators had identified the orientation to drill lateral sections of wells in order to maximize fracture propagation and most wells would be drilled according to this standard [52, 91]. Stage count is likely to be strongly predicted by the combination of lateral length, and total water, so it would be redundant to include along with these other variables [7, 35].

The dependent variable considered in this analysis was the volume of oil produced by a well during the first year. A well’s production rate typically peaks sometime within the first few months and then begins to decline. The first year of production is thus indicative of both the height of this peak and how rapidly production declines, making it a good predictor for the estimated ultimate recovery (EUR), or total volume produced by a well in its decades-long operating lifetime [69, 61]. It is also a good proxy for the economic value of a well because in addition to production in later years being substantially lower, future revenue is subject to a discounting rate and therefore less valuable to investors and oil companies. The reliance on first year production as a metric for productivity also makes it possible to neglect the impact of well interference, in which wells drilled too close to each other end up competing for the same reservoir, thereby reducing production. This behavior is still not well understood, but it is likely to have a negative impact on well production (and economics) in many densely-drilled fields, particularly after wells have been produced for many years and pressure depletion leads to production from reservoir farther away from the well-bore [45, 84].

2.2.2 Multiple linear regression

Three multiple linear regression models were considered, including a completely nonspatial (NS) model, and models controlling for spatial heterogeneity through county-level fixed effects (FE) and surface trend analysis (STA). For the n observed wells, each model incorporated \mathbf{Y} , an n -by-1 vector of the natural logarithm of first year production and an n -by- $(r+1)$ design matrix \mathbf{X} of predictor variables in which each row corresponds to a well and each

column to a variable (with an additional column of ones for the intercept). The natural logarithm was used for production volume both because this is common practice for regression with variables that are positively skewed and by definition nonnegative, and because there is in fact a generally lognormal distribution of productivity in shale and tight resource basins [77]. Although \mathbf{X} differs between these models, they all assume a linear relationship between Y and the predictor variables x_1, \dots, x_r , which can be represented as

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \sigma^2 \mathbf{I}_n) \end{aligned} \tag{2.1}$$

where $\boldsymbol{\beta}$ is the $(j+1)$ -by-1 vector of slopes and intercept for this relationship and $\boldsymbol{\epsilon}$ is the n -by-1 vector of residuals, or error terms. Error terms in this model are assumed to be normally distributed with variance σ^2 and uncorrelated with each other (\mathbf{I}_n is an n -by- n identity matrix). For each model, the sample population (\mathbf{X}, \mathbf{Y}) was used to estimate the regression coefficients $(\beta_0, \beta_1, \dots, \beta_r)$ in order to quantify the strength of the relationships and estimate $E[\mathbf{Y}|\mathbf{X}]$ for forecasting purposes. This was carried out using ordinary least squares in R statistical software, in which $\hat{\boldsymbol{\beta}}$ is the estimate for $\boldsymbol{\beta}$ that minimizes the residual sum of squares, calculated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \tag{2.2}$$

For the NS model, $r = 3$ and the only predictor variables were lateral length (in kft), water volume (in MMgal), and proppant mass (in MMLb)². This model does not control for drilling location in any way. The following two models include these same three variables but include additional predictor variables to control for location.

In the FE model, indicator variables were included in \mathbf{X} for nine of the ten counties present in the data (an indicator for the tenth county would be redundant) and to identify if the target formation was the Middle Bakken (zero if target was the Three Forks). This county-level fixed effects model, which has been used by the EIA to estimate trends in well

²The units adopted are based on oilfield reporting and are also intended to create a well-scaled matrix to avoid numerical issues.

productivity, assumes homogeneity within each county and no relationship between adjacent counties [101]. Any deviation from the mean of a county is attributed to the technology parameters in the model.

For the STA model, a polynomial surface trend was used to account for the large-scale spatial variation in productivity. \mathbf{X} included second-order polynomial terms for well surface location, measured in 10^4 meters east and north from a central point. These predictor variables were *easting*, *northing*, *easting*², *northing*², *easting* \times *northing*. Although other polynomial orders could be chosen, this model was justified *a priori* based on knowledge that the field contains a centrally-located sweet spot. Other polynomial orders explored did not substantially alter the results and are not discussed here. Again a target formation indicator variable was also included to distinguish between the Middle Bakken and Three Forks.

2.2.3 Omitted variable bias and spatial autocorrelation

An important concern in regression modeling is whether $\hat{\beta}$ is an unbiased estimate of β . Bias may result if there are important omitted variables which are correlated with the dependent variable and one or more independent variables. When this occurs, ordinary least squares will compensate by over or under-estimating the strength of relationships represented by coefficients in $\hat{\beta}$. To see this, consider the case where the multiple linear regression model in Eq. (2.1) for productivity should in fact be

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\theta + \epsilon \quad (2.3)$$

where \mathbf{Z} are important omitted variables and θ are the associated relationships to the dependent variable. The parameter estimation in Eq. (2.2) then becomes

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\beta + \mathbf{Z}\theta + \epsilon) \quad (2.4)$$

and the conditional expectation of this estimate is

$$\mathbb{E}[\hat{\beta}|\mathbf{X}] = \beta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{E}[\mathbf{X}^T \mathbf{Z}|\mathbf{X}]\theta \quad (2.5)$$

The second term in Eq. (2.5) introduces bias if $\theta \neq 0$ (there is some relationship between \mathbf{Z} and \mathbf{Y}) and $\text{Cov}(\mathbf{X}, \mathbf{Z}) \neq 0$ (there is correlation between variables in \mathbf{X} and \mathbf{Z}) [64].

Many spatially heterogeneous geological variables, such as permeability and porosity, are influential on well productivity but lack well-level measurements [54, 69], so there are in fact omitted variables \mathbf{Z} with $\theta \neq 0$. Bias will only be introduced if $\text{Cov}(\mathbf{X}, \mathbf{Z}) \neq 0$ though, so if the variations in design parameters in \mathbf{X} were assigned randomly there would be no expected correlation. However, the tendency of operating companies to high-grade activity toward sweet-spots with more favorable geology, alongside improvements to design parameters means that there may be bias resulting from this confounding. In other words, because design choices in \mathbf{X} are improving over time alongside drilling location choices yielding more favorable omitted variables in \mathbf{Z} , changes in \mathbf{X} end up looking more important than they actually are.

As a result of the spatially varying physical processes depositing and transforming a reservoir over time, geological properties tend to vary spatially with a high degree of continuity. In the absence of reliable measurements, location can be used as a proxy for unknown geological properties and the amount of bias introduced by $\text{Cov}(\mathbf{X}, \mathbf{Z})$ depends on how adequately spatial trends in productivity have been controlled for in the regression models. The NS model is likely to have a greater amount of bias than the FE and STA models since these have some controls in place for location. However, given the different scales of heterogeneity in these formations, some reservoir properties may be more localized in nature and vary over length scales too granular for the FE or STA models to control for. For example, operators may focus drilling efforts on a small area with high productivity due to abundant natural fractures or learn to avoid a specific area found to have exceptionally low rock porosity. These spatial patterns may impact multiple locations but will not be reflected in larger-scale productivity trends spanning different areas of the basin.

Spatial dependence of well productivity which has not been controlled for by other variables will appear as spatial autocorrelation of model residuals, in which similar error values have a tendency to be clustered together resulting in a higher degree of correlation between residuals that are near than those far apart [26]. This is an important consideration when conducting multivariate or univariate analysis on data that is observed at different spatial

locations [97]. A standard approach for measuring the amount of spatial autocorrelation in regression residuals is with Moran’s Index I ,

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (\epsilon_i - \bar{\epsilon})(\epsilon_j - \bar{\epsilon})}{\sum_i (\epsilon_i - \bar{\epsilon})^2} \quad (2.6)$$

which requires specification of an n -by- n spatial weights matrix \mathbf{W} (with elements w_{ij}) defining the level of dependence or “connectedness” between data points based on their proximity [64, 25]. There are different ways of constructing \mathbf{W} , depending on the nature of the data, expected spatial phenomena, and modeling task, but it is often chosen to be sparse, with nonzero weights applied only for neighbors within some threshold distance or within the k nearest neighbors of points. In this dataset, wells are irregularly spaced, so it is more appropriate to use k nearest neighbors than a distance threshold [11]. An inverse distance weighting was used for the spatial weights matrix, which is a common scheme to account for rapid spatial decay in autocorrelation of observations. Inverse distance weighting also means that the exact choice k of nearest neighbors to include in \mathbf{W} is relatively unimportant since more distant points have very small weights. A value of $k = 50$ was chosen in accordance with the principle that it is preferable to underspecify rather than overspecify this linkage, but varying this value had little impact on results [31, 26, 25]. Values in \mathbf{W} were also row-normalized, so that each row sums to one.

The possible values for Moran’s I range from -1, which is highly dispersed (negative autocorrelation), to 1, which is highly clustered (positive autocorrelation), with values near zero indicating that regression residuals appear spatially random and uncorrelated. It is important to recognize that the assumptions used for the structure of \mathbf{W} are an inexact representation of actual autocorrelation and will impact the measure of Moran’s I . However, for a given dataset and choice of \mathbf{W} , it is a useful metric for comparing the effectiveness of different regression models at controlling for spatial dependence at the scale represented in \mathbf{W} . The modeling of the spatial weights matrix and measurement of Moran’s I was performed with SPDEP in R.

2.2.4 Regression models accounting for spatial autocorrelation

Two additional regression formulations were considered—a spatial error model (SEM) and a regression-kriging (RK) approach—which incorporate spatial autocorrelation into estimates and have not been previously applied to this topic in literature. SEM was developed in the spatial econometrics field but has also been applied to urban energy consumption and ecology [4, 64, 97, 25, 26]. RK, which is also often referred to as universal kriging, kriging with external drift, or feasible generalized least squares (with only slight differences in formulation), has been applied to a variety of earth science applications, such as soil properties [80, 39, 33]. These models are based on changing the assumptions in Eq. (2.1) so that residuals are recognized to be correlated according to a variance-covariance matrix $\mathbf{\Omega}$, as in

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(0, \mathbf{\Omega}) \end{aligned} \tag{2.7}$$

An unbiased estimator $\hat{\boldsymbol{\beta}}$ can be obtained under these assumptions by using $\mathbf{\Omega}$ to remove correlation between observations, which is referred to as generalized least squares, shown in Eq. (2.8)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Omega}^{-1} \mathbf{Y} \tag{2.8}$$

However, the true $\mathbf{\Omega}$ matrix is unknown, so the correlation of residuals must be estimated alongside estimates of $\boldsymbol{\beta}$. SEM and RK are two common approaches for this.

The spatial error model (SEM) breaks up the error term into a correlated component \mathbf{e} and an uncorrelated error component \mathbf{u} , as

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \rho \mathbf{W}\mathbf{e} + \mathbf{u} \\ \mathbf{u} &\sim N(0, \sigma^2 \mathbf{I}_n) \end{aligned} \tag{2.9}$$

where the correlated error term is weighted by the pre-defined spatial weights matrix \mathbf{W}

(here assumed to be inverse distance weighting, as discussed in Section 2.2.3) and a single parameter ρ to adjust the scale of autocorrelation [25, 4]. Estimates of ρ and σ^2 can be obtained by optimizing the log-likelihood, leading to an estimate of Ω as

$$\hat{\Omega} = \hat{\sigma}^2[(\mathbf{I}_n - \hat{\rho}\mathbf{W})^T(\mathbf{I}_n - \hat{\rho}\mathbf{W})]^{-1} \quad (2.10)$$

which can then be used in Eq. (2.8) to estimate β . This approach is discussed in detail in [64] and was implemented here using the SPDEP package in R. A Monte Carlo approximation of the log-determinant was also used to facilitate the computation, since this has been found to result in very little loss of accuracy [64].

Regression-kriging (RK) similarly breaks the error term up into a correlated and uncorrelated component, but an iterative process is used to achieve this. RK can be defined as

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \lambda\mathbf{e} + \mathbf{u} \\ \mathbf{u} &\sim N(0, \sigma^2\mathbf{I}_n) \end{aligned} \quad (2.11)$$

where λ is referred to as the kriging weights matrix and $\mathbf{X}\beta$ is often referred to as the “drift” or mean component [39]. Initially, β is estimated using ordinary least squares (Eq. 2.2) and ignoring the autocorrelation of error, leading to an estimate of the error as

$$\hat{\epsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta} \quad (2.12)$$

Spatial autocorrelation within $\hat{\epsilon}$ is then measured in terms of semi-variance using

$$\gamma(h) = \frac{1}{2N} \sum_i^n \sum_{j; h_{ij}=h} (\epsilon_i - \epsilon_j)^2 \quad (2.13)$$

which sums over the N pairs of residuals $(\hat{\epsilon}_i, \hat{\epsilon}_j)$ that are spaced h apart (within some tolerance). This is calculated over various values of h in order to develop an empirical semi-variogram describing the nature of autocorrelation in $\hat{\epsilon}$. Parameters of a theoretical semi-variogram model are then estimated based on this empirical semi-variogram using ordinary

least squares. The choice of semi-variogram model used here was the exponential model, defined as

$$\gamma(h) = \tau^2 + \sigma^2[1 - \exp(-h/\phi)] \quad (2.14)$$

where τ^2 , σ^2 , and ϕ are model parameters [39]. This semi-variogram model is frequently used in geostatistics for geological properties, and was found to provide a good match to residual autocorrelation in this analysis [39, 33].

The kriging weights³ matrix $\boldsymbol{\lambda}$ is obtained from this semi-variogram model (along with the spatial configuration of observation and prediction locations) in order to minimize error variance as described in [33]. The autocorrelated error term can now be estimated as

$$\hat{\boldsymbol{e}} = \boldsymbol{\lambda}\boldsymbol{\hat{e}} \quad (2.15)$$

and based on this, a new estimator $\hat{\boldsymbol{\beta}}$ can be obtained with the autocorrelated error component removed, as in

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{Y} - \boldsymbol{\lambda}\hat{\boldsymbol{e}}) \quad (2.16)$$

Using this new $\hat{\boldsymbol{\beta}}$, the procedure described in Eq. (2.12) to Eq. (2.16) can then be repeated until parameters in the semi-variogram model converge. Here, the process was repeated until all three parameters in Eq. (2.14) changed less than 0.1% from the previous estimate, which required three iterations.

RK was carried out using the geoR package in R. Both SEM and RK included the same design matrix X as STA in order to remove large-scale trends. A summary of the five regression models in this study can be found in Table 2.1.

2.2.5 Evaluating the results

The regression model dependent variable Y is related to the first year production volume Q (measured in Mbbl) by $Y = \log(Q)$. Exponentiation of predictions of Y to the original scale

³Simple kriging weights are used (as opposed to ordinary kriging weights) since \boldsymbol{X} has already de-trended the data.

	NS	FE	STA	SEM	RK
Form	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{e} + \mathbf{u}$	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \lambda\mathbf{e} + \mathbf{u}$
Technology variables in \mathbf{X}	Lateral length, water volume, proppant mass	Lateral length, water volume, proppant mass	Lateral length, water volume, proppant mass	Lateral length, water volume, proppant mass	Lateral length, water volume, proppant mass
Additional variables in \mathbf{X} to control for location	N/A	County indicators, formation indicator	Second order polynomial of coordinates, formation indicator	Second order polynomial of coordinates, formation indicator	Second order polynomial of coordinates, formation indicator
Fitted parameters to control for spatial auto-correlation	N/A	N/A	N/A	ρ	τ^2, σ^2, ϕ
Decay of spatial auto-correlation assumed	N/A	N/A	N/A	Inverse distance weighting, first 50 neighbors only	Exponential

Table 2.1: Summary of the regression models used.

will result in transformation bias though, and it is theoretically necessary to use a confluent hypergeometric function for all back transformations to the original scale, as discussed in [92]. However, a suitable approximation is to estimate σ^2 as the mean squared error and use the approximation

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^n ((Y_i - Y_i^*)^2) \quad (2.17)$$

$$Q_i = \exp(Y_i + \hat{\sigma}^2/2) \quad i = 1, \dots, n$$

where Y_i^* is the prediction for Y_i . For the results in this chapter, there was no discernible difference between this approximation and the hypergeometric transformation. It is beyond

the scope of this chapter to discuss when bias would be introduced by using the approximation in Eq. (2.17) rather than the hypergeometric function and this appears to be an area requiring further research.

In order to compare the prediction accuracy of models to each other and avoid overfitting of models to the data, a 10-fold cross validation (CV) was used. This approach involves randomly dividing the data into 10 equal sized subsets. One at a time, a subset was held out while the model was fit to the remaining 90% of the data. Accuracy of the model was then calculated based on predictions for the data that had been held out. As an additional validation of the models, a hindcasting approach was used, in which only wells from the first half of the overall time period were used for fitting, and predictions were then made for all wells from the second half of the overall time period.

As a standard measure of prediction accuracy, the mean absolute scaled error (MASE) was calculated using,

$$\text{MASE} = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i - Y_i^*|}{|Y_i - \bar{Y}|} \quad (2.18)$$

where the error of each prediction is scaled by a baseline prediction using simply the mean of available data \bar{Y} and these scaled errors are averaged. As such, values less than one are an improvement over predictions using the mean of the data and zero represents a perfect prediction. This statistic is a transparent way of understanding a model's success at making actual predictions [43].

2.3 Results and discussion

2.3.1 Regression model estimates

A summary of selected metrics of performance for each regression model is shown in Table 2.2 (additional metrics of model performance are included in supplementary table S1). A disadvantage of the RK approach is the computational time, which is substantially higher than other approaches. This is due in part to the lack of sparsity in the weights matrix λ and the need to iterate to convergence. In this analysis, after 3 iterations of the algorithm

semi-variogram parameters changed less than 0.1%.

	NS	FE	STA	SEM	RK
Time to run (s)	0.0041	0.00804	0.00828	3.7	478.54
MASE	0.938	0.871	0.815	0.662	0.532
10-fold CV MASE	0.938	0.873	0.816	0.669	0.62
Moran's I (\mathbf{W})	0.512	0.443	0.403	-0.00895	-2.26E-04
Moran's I ($\boldsymbol{\lambda}$)	0.548	0.482	0.444	0.245	0.102

Table 2.2: Comparison of performance for the regression models. Moran's I was calculated with both the inverse distance weighted matrix \mathbf{W} and the kriging weights matrix $\boldsymbol{\lambda}$.

Prediction accuracy improves with increasing spatial fidelity of the models, with fairly poor prediction accuracy by the NS model, which accounts for no spatial heterogeneity, and the best performance by SEM and RK, which explicitly model autocorrelation. The FE model is also relatively inaccurate, suggesting that despite differences in average productivity between counties, spatial heterogeneity within counties is important and missed by this approach. The improvement between STA and SEM is due to the ability to make use of information on local patterns provided by spatial autocorrelation and the further improvement with RK is due to the use of a more flexible and empirical estimate of the variance-covariance matrix $\boldsymbol{\Omega}$. Model predictions from 10-fold CV are compared to actual values in Fig. 2-2 for four of the models.

Over-fitting of models can be identified when cross-validation and non-cross-validation predictions differ in accuracy. This was only an issue with RK, where there was a drop in prediction accuracy under CV. RK still provides the best predictions in 10-fold CV though so this is more likely indicative of a plateauing in accuracy that can be achieved given the noisiness of the data being modeled.

Spatial autocorrelation of regression residuals (Moran's I) is substantially reduced by using the SEM and RK models, which control for this feature. Moran's I was measured with both the *a priori* inverse distance weighted matrix \mathbf{W} and the kriging weights matrix $\boldsymbol{\lambda}$ described in Section 2.2.3. The difference in these weights matrices, illustrated in Fig. 2-3 leads to differences in estimates by SEM and RK. From this it can be seen that SEM places a greater weight on the nearest neighbor but less weight on more distant neighbors. Although SEM helps to reduce spatial autocorrelation by using an assumed inverse distance weighting

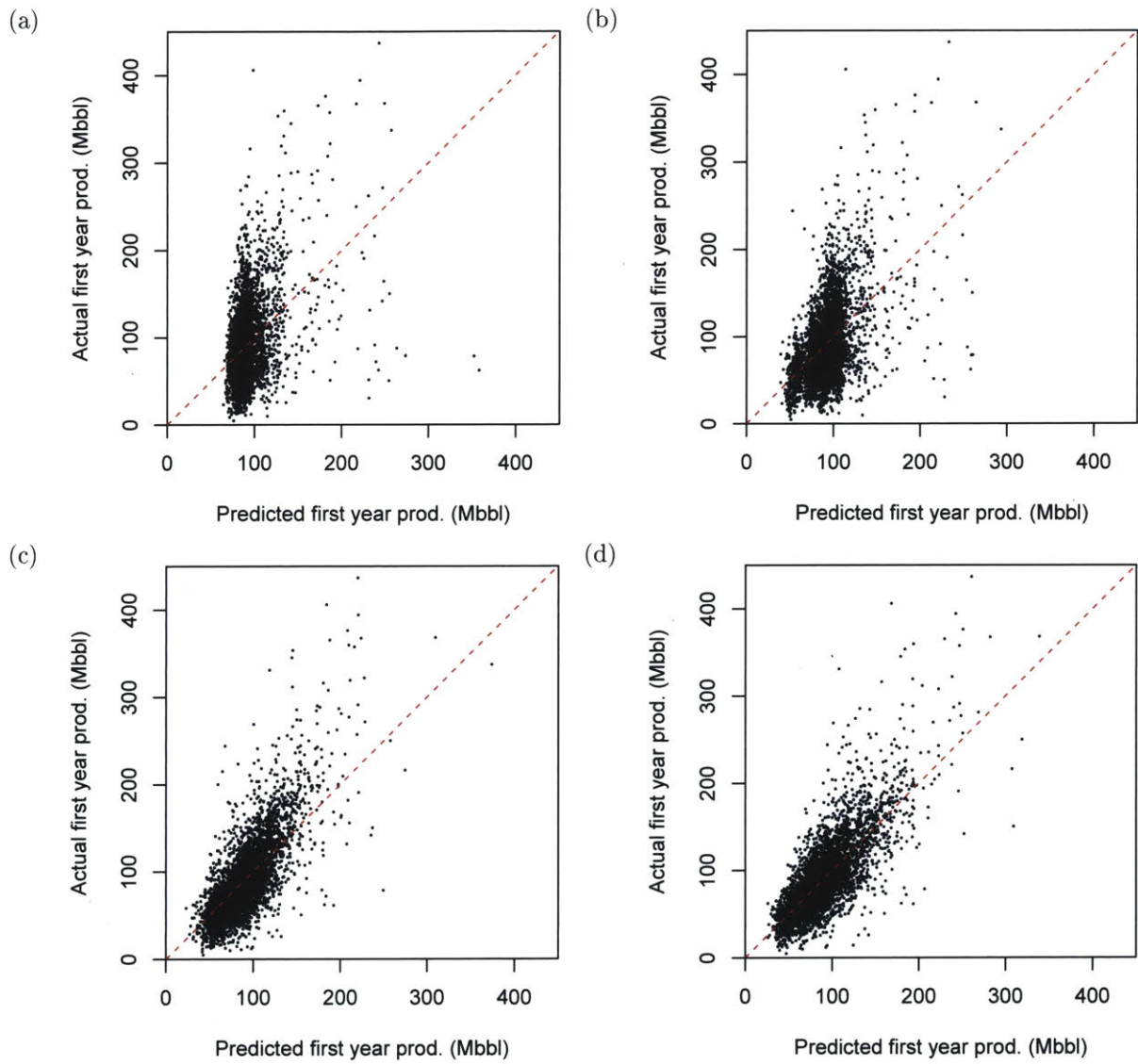


Figure 2-2: Predicted (with 10-fold cross-validation) and actual values for four of the regression models. The dashed line indicates a perfect prediction. (a) NS, (b) FE, (c) SEM, (d) RK.

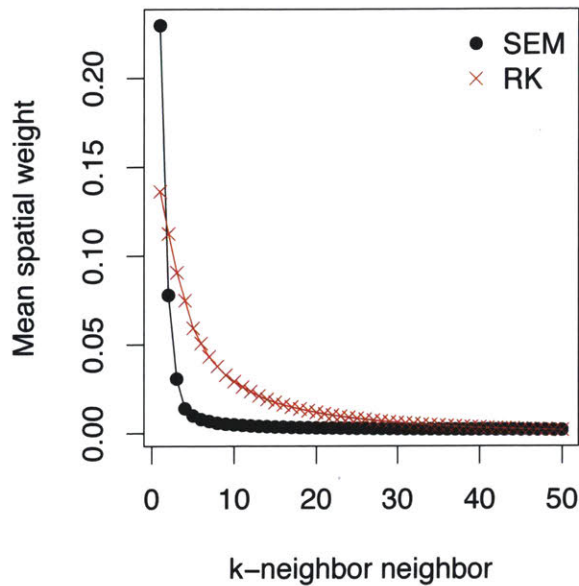


Figure 2-3: Comparison of average spatial weights used by SEM (inverse distance weighted) and RK (exponential semi-variogram model) for each of the nearest 50 neighbors.

scheme, this approach is inexact and the more robust estimate of spatial autocorrelation achieved with RK appears to be necessary in this context.

Examining the coefficient estimates for lateral length, water volume, and proppant mass from each regression model reveals that spatial resolution has an impact on how increases in productivity are attributed to these design choices. As the spatial fidelity of models increases, lateral length is inferred to have a greater impact and proppant mass a lesser impact on production volumes. This is a critical insight since lateral length of wells has leveled out in the Williston basin at around 9,200 ft, but proppant mass and water volumes are expected to rise to nearly double the levels of 2014 by 2018. This shift is making wells more environmentally and economically costly but may fail to yield the magnitude of productivity increase expected by forecasters relying on regression models biased by spatial autocorrelation.

2.3.2 Forecasting applications

An important application of the regression models is as a predictive tool, particularly for mean trends in well productivity, as introduced in Fig. 2-1. Each model was used to make

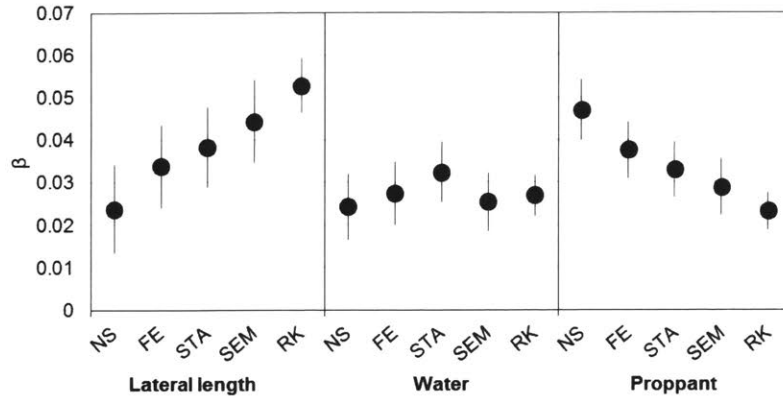


Figure 2-4: Regression parameter estimates for lateral length, water volume, and proppant mass by each model with a 95% confidence interval.

predictions of first year production based on the variables from \mathbf{X} actually used in wells and these results were averaged within each quarter in order to estimate this trend in mean well productivity over time. The results of this are shown in Fig. 2-5 compared to the actual mean trend in new well productivity over the time period studied. Despite relying on only a few variables, even the simpler linear regression models are well tuned to reflect the trend of increasing productivity over time. This is not surprising however, since the models were fit to the same data they are being compared to here. As a more robust demonstration of productivity trend forecasting, the regression models were trained with data only from the first half of the time period and then used to predict production for all wells. The resulting predicted trends are shown in Fig. 2-6, and there is still a striking agreement between predicted and actual trends. This suggests that the regression models have a capacity to make future predictions of well productivity given knowledge of where drilling will take place and the technology that will be used.

Although the mean productivity trends predicted by all five models are similar for the time period analyzed, the differences in regression coefficient estimates (Fig. 2-4) lead to substantially different forecasts when near-term future trends in hydraulic fracturing design are considered. The EIA and IHS, an energy data consultant, have estimated that by 2018, the average proppant mass used for wells in the Williston basin will rise to 7.7 MMlb (it was 5 MMlb in 2015) with water volumes rising proportionally [103]. Using these technology variables for wells at the locations drilled in the first half of 2015, stark differences become

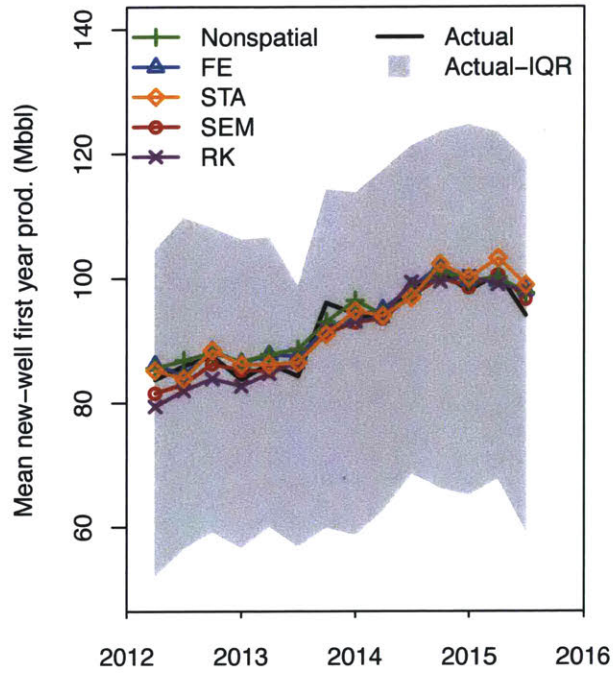


Figure 2-5: The mean and interquartile range (IQR) of actual well productivity in each fiscal quarter over the time period studied and the mean prediction for each model.

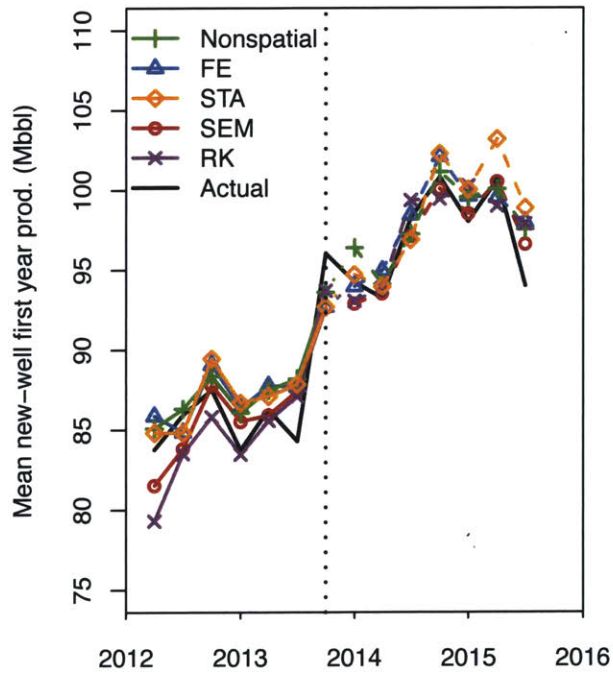


Figure 2-6: The mean of actual well productivity in each fiscal quarter over the time period studied and the mean prediction for each model. The models are trained only with data from wells to the left of the dotted line.

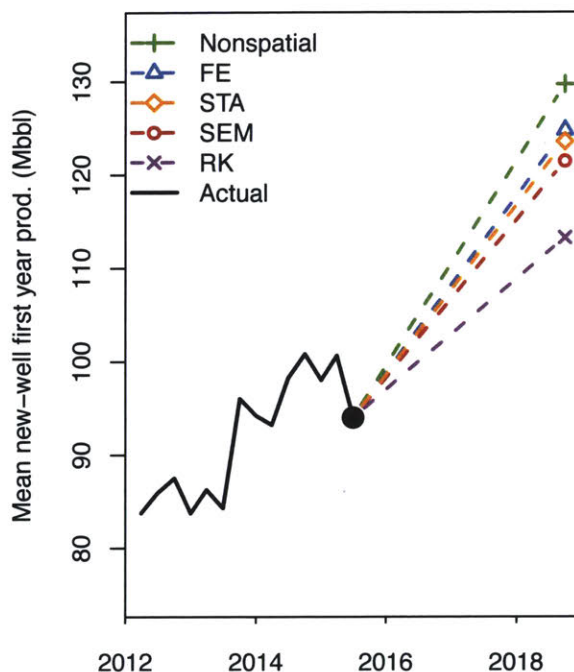


Figure 2-7: Forecast of mean first year production for wells drilled in the first half of 2015, but using design parameters expected for 2018 (based on EIA-IHS report [103]). These estimates do not reflect any potential changes in drilling locations. The distributions for each prediction are shown in Table 2.3.

clear between model predictions, as shown in Fig. 2-7. Operating companies have already experimented with hydraulic fracturing designs of this magnitude and larger in the Williston basin and they are increasingly shifting designs in this direction, in part based on statistical analysis of data suggesting substantial scope for improved productivity and economics of wells. Although it is hard to know what analysis techniques are being used internally by operating companies, it is likely that in some cases these design decisions are being made based on regression models failing to control for the endogeneity introduced by high-grading of drilling activity to the most productive locations. Wells designed based on those models will overshoot economically optimal quantities of water and proppant as a result of this conflation.

Although the forecasts in Fig. 2-7 reflected no change in drilling locations from 2015, more granular forecasts can be developed based on where wells will be drilled in the future. Fig. 2-8 shows a heat map of the RK predicted first year well production for all locations within

2018 forecast of first year prod. (Mbbbl)				
Model	P25	Median	Mean	P75
NS	90.7	125.3	129.7	160.4
FE	86.4	121.3	124.8	154.3
STA	85.8	119.5	123.6	153.9
SEM	85.0	117.1	121.5	147.9
RK	76.5	109.9	113.3	139.7

Table 2.3: The distribution of predicted productivity for wells drilled in the first half of 2015 with technology parameters expected in 2018 [103].

the Bakken region of North Dakota, assuming the average technology design parameters expected by the EIA for 2018 [103]. Others have developed interpolations or gridded heat maps based on well productivity in oil and gas fields, but the use of RK here improves on this substantially by controlling for differences in technology and presenting spatial variability in productivity with a standardized well design [45, 47, 112, 21]. These results could be further applied to develop more robust supply curves for different economic scenarios or to optimize design parameters based on location.

2.3.3 Dis-aggregating the productivity trend

One approach to understanding the relative role played by a variable in a multiple regression model is by holding constant the other parameters in the model while allowing the selected parameter to vary according to the data [8]. This technique was used to isolate the change in well productivity associated with changes in drilling location from changes in lateral length, proppant mass, and water volume over the time period considered. For each well, a prediction was made based on its location, with other parameters held constant. The result of this approach for the four spatially-resolved regression models is shown in Fig. 2-9. There are large differences throughout the time series, particularly between FE and STA compared to SEM and RK, in the amounts of productivity improvement attributed to changes in location.

This same approach was carried out to isolate the impact of variation in each design parameter on well productivity. For each design parameter, location and the other two

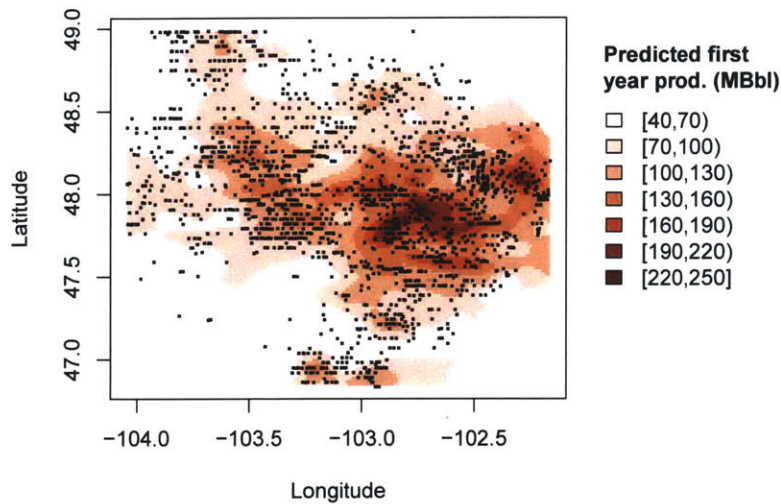


Figure 2-8: Predictions of first year production for locations within the Bakken formation in North Dakota using design parameters expected for 2018 [103]. The markers represent wells in the dataset.

design parameters were held constant while the selected design parameter was allowed to vary according to data. The first quarter of 2015 was compared to the first quarter of 2012 to determine the relative impact of the more productive choices being made after a three year span. The impacts on productivity associated with location and technology were found to be roughly even by SEM and RK, but the other models differed considerably.

This again highlights the important difference in how these models control for location and the bias that can result when these controls are inadequate. The FE model, which misses sub-county-level sweet spotting is of particular concern since it is used by the EIA to estimate rates of technological learning for their low and reference case forecasts, yet overestimates technology improvement by a factor of two (their high resource case assumes an additional 1% per year “accelerating” of technological learning). Despite having controlled for differences in location through county-level fixed effects, this approach still ends up attributing nearly all of the effects of sweet-spotting to technology. This bias has the potential to lead to unrealistic forecasts of drastically improved well productivity and economics, which are in fact only an extrapolation of transient effects from the drilling of wells in better locations.

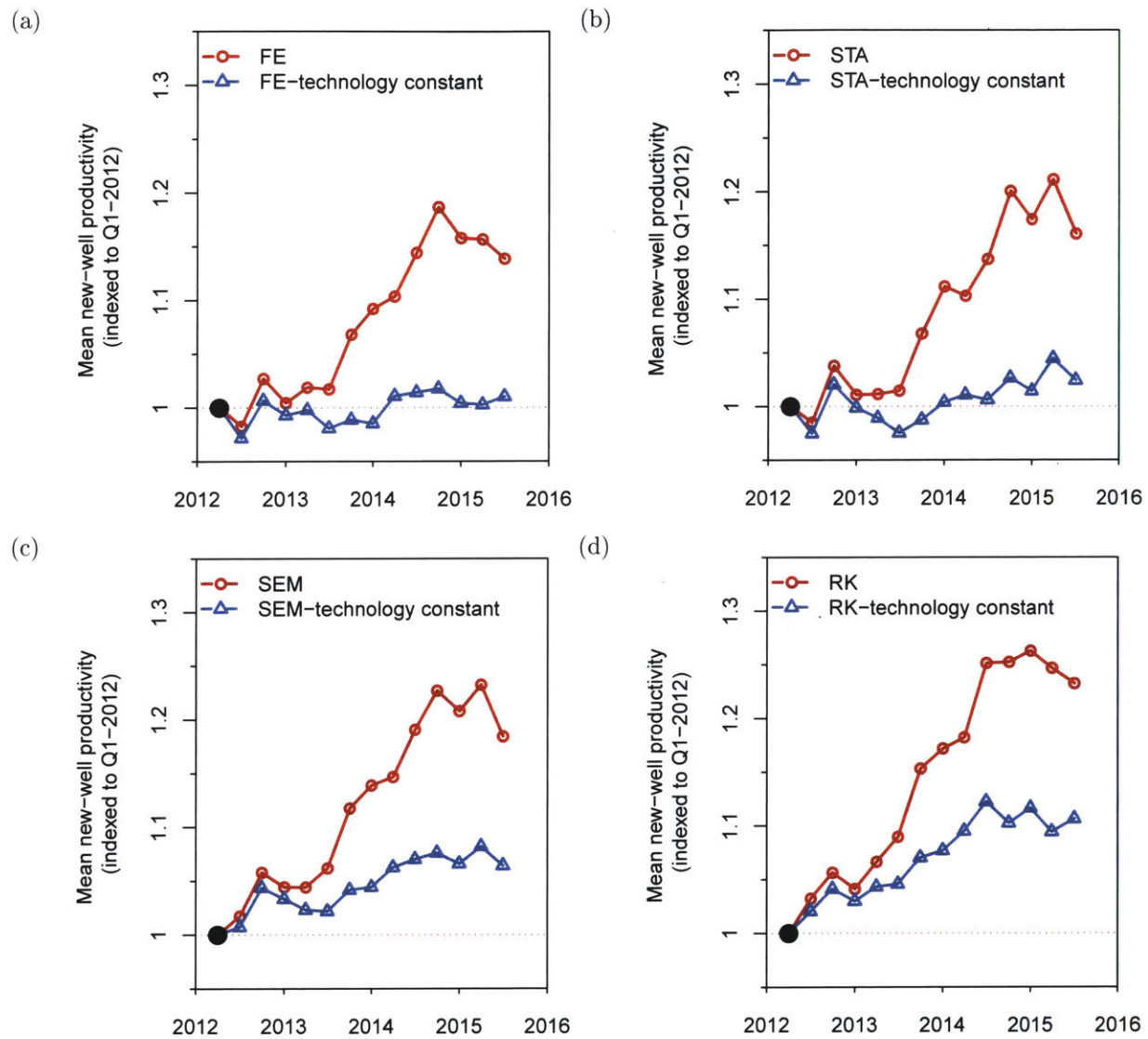


Figure 2-9: Comparison of predicted mean well productivity to predictions with technology held constant to reflect the impact of location. (a) FE, (b) STA (c) SEM, (d) RK.

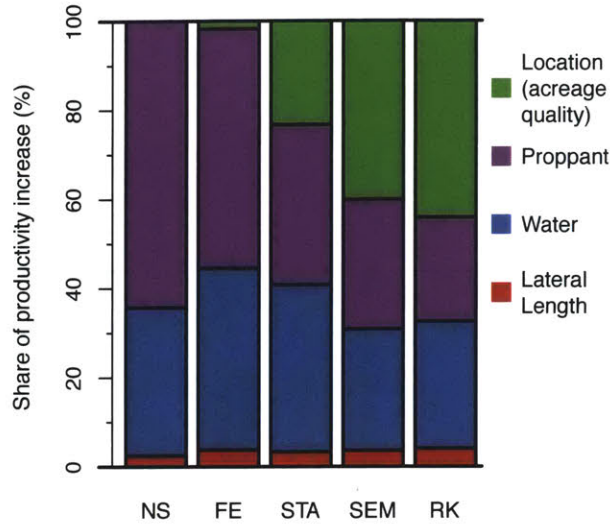


Figure 2-10: Breakdown of the relative influence of factors on the productivity improvement between Q1-2012 and Q1-2015.

2.4 Concluding remarks

This chapter analyzed a large contemporary tight oil well dataset from the Williston Basin in order to quantify the extent to which improvements in well productivity have been associated with the scaling up of well and hydraulic fracturing designs as opposed to changes in development location. Using five regression models of increasing spatial resolution it was demonstrated that the results of analysis aiming to answer this question are highly sensitive to the statistical techniques used. Unless the chosen modeling approach incorporates high-resolution spatial dependence a significant bias can be introduced leading to overestimates of the relative impact of technology on well productivity improvement. This is in line with statistical theory about omitted variable bias, but has not previously been recognized for this important energy application. Of particular concern here is the EIA’s reliance on a county-level fixed effects approach to develop recent national production projections since this model was found to overestimate the role of technology by a factor of two. This may result in projections that are overly optimistic regarding the future potential of the nation’s unconventional resources.

Results from regression-kriging and a spatial error model—regression approaches that incorporate high-resolution spatial dependence through spatial autocorrelation—concluded

that the impact of technology on well productivity has been roughly equivalent to the impact from the high-grading of drilling locations in the Williston Basin. Crucially, this result means that if future development requires expansion into acreage outside of existing geologic sweet spots, half of the well productivity gains achieved in recent years will not be transferable to these other parts of the basin. Although this specific conclusion cannot yet be directly extended to the broader set of tight oil and shale gas regions under development today, very similar trends in terms of productivity gains, technology changes and acreage high-grading have taken place over the recent past and so the conclusions from the Williston analysis presented here are likely to be more broadly applicable.

Chapter 3

The fundamental ambiguity in mechanistic-statistical production forecasting

Decline curve analysis (DCA)—the extrapolation of a production curve model fitted to a well’s past production—remains the standard approach for forecasting unconventional oil and gas production. A scaling curve based on a fractured shale gas reservoir model was recently proposed as a way of connecting this approach with underlying physics but as this chapter shows, it actually generates worse predictions than the traditional non-physical modified Arps curve. DCA is fundamentally an ill-posed inverse problem with the defining characteristic of model sloppiness, or parameter correlation. Today’s unconventional resource forecasts can be substantially improved by using information from offset wells to reduce ill-posedness through Tikhonov regularization. This versatile approach nearly matches a deep neural network approach introduced here, which has practical limitations but offers a model-neutral benchmark of achievable extrapolation accuracy. There is a natural connection between regularization and a Bayesian formulation which is also highlighted. This chapter evaluates long-term forecasting accuracy for these techniques using historic production data from 4457 Barnett shale wells, and reveals that the overlooked step of regularization is more critical than choice of model.

3.1 Introduction

Effectively planning for and managing the development of unconventional oil and gas resources requires accurate production forecasts which remain elusive due to a limited understanding of the complex physical processes behind production [61, 68, 95]. Many wells under-perform compared to operators’ projections and energy policy decisions are hampered by the immense uncertainty in the long-term productivity of new wells [3, 83, 104, 48].

Rigorous physics-based models, such as numerical reservoir simulators, are rarely used to forecast unconventional oil and gas production due to the associated data and modeling costs as well as the challenge of adequately representing nano-scale flow and fracture properties [61, 68]. Instead, it is standard to rely on decline curve analysis (DCA), in which a production curve model with a small number of parameters is matched to historical production rates using a least squares fit and the resulting curve is extrapolated into the future [61]. The most widely used DCA model is the Arps curve, which builds on the hyperbolic decline relationship,

$$Q(t) = \frac{Q_i}{(1 + bD_it)^{\frac{1}{b}}}, \quad (3.1)$$

originally introduced in 1945 [5], in which production rate $Q(t)$ at time t is determined by the parameters Q_i , b , and D_i . In order to avoid unrealistically high late-life projections for unconventional wells, a modified Arps curve is used in which the hyperbolic model in Equation 3.1 is switched to a fixed exponential trend once production decline slows to some threshold rate, such as 10% annually [95, 105]. Although widely used, DCA with modified Arps is a heuristic approach and lacks a physical basis with unconventional wells.

In order to better understand shale gas production decline behavior and put DCA on firmer theoretical ground, Patzek et al. developed a scaling curve model (SCM) based on one-dimensional gas flow into planar fractures [84]. Remarkably, this physics-based model could be reduced to two effective “scaling” parameters, \mathcal{M} and τ , and a recovery factor curve, or forward model $F(\cdot)$, to describe any Barnett shale well’s cumulative production $Q(t)$ at time t , as in

$$Q(t) = \mathcal{M}F(t/\tau). \quad (3.2)$$

Because physical properties behind the parameters \mathcal{M} and τ cannot be measured, Patzek et al. advocated a nonlinear least squares fit to production data, as is standard for DCA. The SCM will be discussed further in Chap. 4.

There has been a proliferation of production model formulas in recent years—ranging from purely empirical, like the modified Arps curve, to those that build upon physical theory, such as the SCM [61, 84]. There is ongoing debate about which model is most appropriate for unconventional wells, as reviewed in [95]. However, as this chapter shows, there is a glaring shortcoming with the parameter estimation process of DCA more generally, which is evident in both these representative DCA models and actually worse with the SCM.

DCA forecasting as currently implemented is unreliable due to potentially non-unique parameter estimates, particularly when the production history is short. This has also been described as a sensitivity in forecasts to initial parameter seeds in the nonlinear least squares algorithm [41]. More saliently, this property of non-unique or non-identifiable parameters makes DCA a classic example of an ill-posed inverse problem which lacks a stable numerical solution [94, 55, 29]. Consequently, DCA should incorporate regularization, as is typically used with this class of problems, but this critical step has surprisingly been overlooked in this context.

Regularization is the process of introducing additional information to reduce the ill-posedness of an inverse problem. In machine learning, it also helps to avoid over-fitting by introducing some bias to reduce the overall variance of predictions [55, 10]. Often, this takes the form of ℓ^2 Tikhonov regularization, where a squared penalty on the distance of parameter estimates θ from some expected parameter value θ_0 is added to the objective function. In statistics, this is often also called ridge regression. For least squares problem, the loss function $L(\mathbf{t}, F(\mathbf{t}, \theta))$ being minimized for observations \mathbf{y} at times \mathbf{t} with forward model predictions $F(\mathbf{t}, \theta)$, becomes

$$L(\mathbf{t}, F(\mathbf{t}, \theta)) = \|\mathbf{y} - F(\mathbf{t}, \theta)\|_2^2 + \lambda \|\theta - \theta_0\|_2^2. \quad (3.3)$$

The ℓ^2 -norm is denoted as $\|\cdot\|_2^2$ and λ is a weight hyperparameter controlling parameter shrinkage toward the mean, which can be tuned to optimize the bias-variance tradeoff and

ensure model generalizability using cross-validation [55, 29].

When forecasting with limited well production data, it is currently common to avoid inversion altogether and rely on a type-well curve as a proxy. This is simply an average of production rates from older wells in the same area [105, 68]. Ad hoc approaches may be used to normalize type-well curves to design metrics (e.g. lateral length) or peak production rate, but this ignores the impact of variations in geology and evolving development practices on production dynamics. This is a particular concern as newer wells tend to be drilled much closer to neighboring wells and have more tightly spaced fracture stages in order to rapidly drain an area [3, 83]. As shown in Chap. 2, geology is exceptionally heterogeneous at even small scales in these basins and it is difficult to isolate the impact of constantly evolving technology on production, making the selection of suitable analogue wells for a type-well curve a highly fraught task.

The accuracy of these forecasting approaches is quantified using the error in predictions of 10-year cumulative production for 4457 wells in the Barnett shale. These projections are made based on the first 6, 12, 24, and 48 months of production. The DCA models considered here—modified Arps and the SCM—are compared to a static county-based type-well curve, as established by the US Energy Information Administration (EIA) [105]. The same data processing and fitting procedure as [84] is followed for the SCM and for modified Arps the methodology laid out in [105] is used. These DCA models are then fitted with ℓ^2 Tikhonov regularization to understand the benefit this provides. A grid-search with 4-fold cross-validation is used to tune the regularization parameters λ and θ_0 , as described in [55, 29]. This consists of four rounds of training and testing, in which λ and θ_0 are adjusted to minimize the root mean squared error (RMSE) within the training data. In each round of cross-validation, the regularization values obtained from fitting the training wells are then used with the corresponding test set of wells, which have been set aside, to measure the regularized models' accuracy at making out-of-sample predictions.

Given the many physical complexities of unconventional oil and gas production, a DCA model with a small number of parameters considerably oversimplifies production dynamics resulting in model mis-specification error. To better understand the magnitude of this error an alternative data-driven extrapolation approach using a deep neural network (DNN) in

the TensorFlow platform is used. This consists of 4 hidden layers of 15 fully-connected neurons (with ReLU activation functions), a regression layer outputting 10-year cumulative production, and an input layer with the number of input neurons determined by the number of production months used in the forecast. As with Tikhonov hyperparameter tuning, 4-fold cross-validation is used to separately train and test the DNN.

3.2 Analysis

Prediction accuracy, measured as RMSE across all four rounds of cross-validation test wells, is shown in Fig. 3-1. Despite its physical derivation, the SCM introduces substantial error compared to the modified Arps model. By the time 48 months of production data are available for fitting, the modified Arps model with regularization nearly matches the reliability of the DNN forecast. The regularized SCM at this point performs slightly worse, but actually is more accurate at 6 months.

One way to better understand the source of ill-posedness in this inverse problem is by examining a related property: the sloppiness of the models themselves. This is an intrinsic property of the models that measures how sensitive they are to parameter changes in certain directions [98]. Put another way, it is how correlated model parameters are with each other. In some “stiff” directions of the parameter space, parameter changes will cause drastic changes in a model’s behavior, while in sloppy directions there is very little influence. This insensitivity leads to parameter estimates that are underidentified since they can vary along this sloppy direction over large ranges of values while giving similar model responses. This is a common property of ill-posed problems and makes regularization even more critical when it is present [29].

Sloppiness is typically measured by examining the curvature of the loss function at a particular parameter combination. As described in [98], this can be accomplished using the approximate Hessian H of the loss function at a particular set of parameters, as in

$$H \approx J^T J \tag{3.4}$$

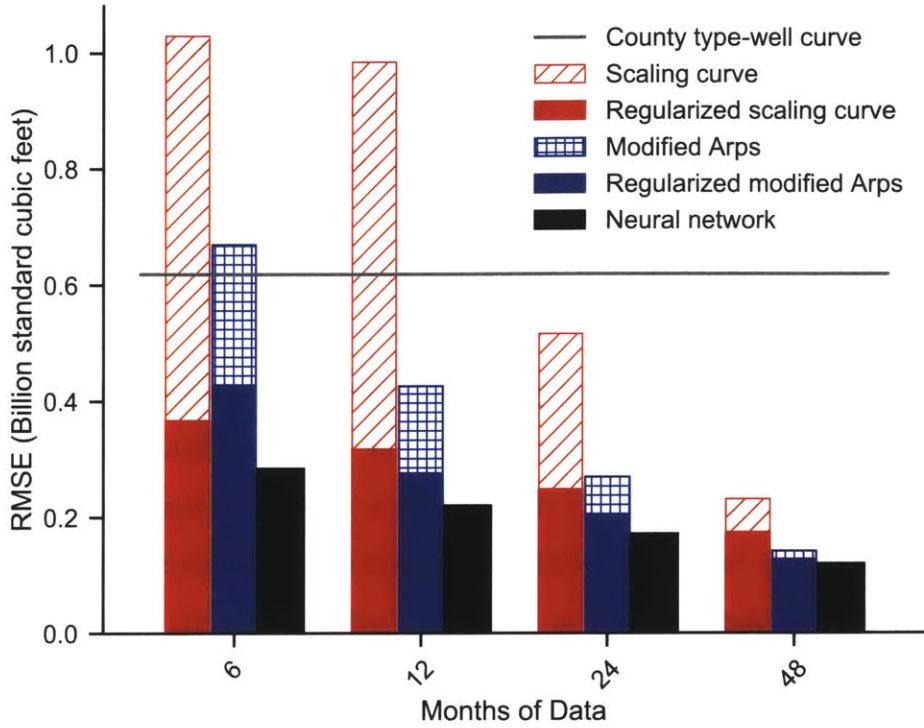


Figure 3-1: A comparison of prediction accuracy for 10-year cumulative production with different approaches and months of production data.

where J is the Jacobian of the residuals

$$r_m = \frac{Q(t_m) - \hat{Q}(t_m)}{\sigma}. \quad (3.5)$$

The Jacobian for the SCM is

$$J = \begin{bmatrix} \frac{\partial r_1}{\partial \mathcal{M}} & \frac{\partial r_1}{\partial \tau} \\ \vdots & \vdots \\ \frac{\partial r_M}{\partial \mathcal{M}} & \frac{\partial r_M}{\partial \tau} \end{bmatrix} \quad (3.6)$$

where

$$\frac{\partial r_m}{\partial \mathcal{M}} = \frac{-F(t_m/\tau)}{\sigma} \quad (3.7)$$

and using the chain rule with $\tilde{t} = t_m/\tau$,

$$\frac{\partial r_m}{\partial \tau} = \frac{\mathcal{M}}{\sigma} \frac{\partial F(t_m/\tau)}{\partial \tilde{t}} \frac{t_m}{\tau^2}. \quad (3.8)$$

Note that since $F(\cdot)$ is a tabulated function, the derivative must be approximated using finite differences. For the Arps curve, the Jacobian is

$$J = \begin{bmatrix} \frac{\partial r_1}{\partial Q_i} & \frac{\partial r_1}{\partial D_i} & \frac{\partial r_1}{\partial b} \\ \vdots & \vdots & \vdots \\ \frac{\partial r_M}{\partial Q_i} & \frac{\partial r_M}{\partial D_i} & \frac{\partial r_M}{\partial b} \end{bmatrix} \quad (3.9)$$

where

$$\frac{\partial r_m}{\partial Q_i} = \frac{1}{\sigma} \frac{-1}{(1 + bD_it_m)^{(1/b)}} \quad (3.10)$$

and again using the chain rule

$$\frac{\partial r_m}{\partial D_i} = \frac{1}{\sigma} Q_i t (bD_it + 1)^{\frac{-(b+1)}{b}} \quad (3.11)$$

and

$$\frac{\partial r_m}{\partial b} = \frac{-1}{\sigma} Q_i (bD_it + 1)^{(-1/b)} \left[\frac{\log(bD_it + 1)}{b^2} - \frac{-D_it}{b(bD_it + 1)} \right]. \quad (3.12)$$

The eigenvalues of the Hessian indicate how sensitive the model is to parameter changes in different directions. Sloppiness is defined as the ratio of the largest (stiff direction) eigenvalue to the smallest (sloppy direction) eigenvalue of this Hessian. In a sloppy model, this ratio is typically at least 10^3 . The calculated eigenvalue ratio is plotted for each model with typical parameter values as a function of the number of months m included in the t observations in Fig. 3.2. This shows that sloppiness is inherent to both of these models and is most extreme when limited production is available. Even after significant data is available for fitting, there is still a large amount of parameter correlation. The unevenness with the SCM is expected since it is a piece-wise tabulated function. It is unsurprising that the ratios are generally higher with the Arps curve since it has an additional parameter.

The behavior that this describes can also be observed directly in Fig. 3.2, which shows the loss function associated with fitting the SCM to the first 12 months of production data for a well without and with regularization. In this context, the sloppiness can be understood as the fundamental ambiguity between a large SRV being drained slowly (large \mathcal{M} and large τ) and a small SRV being drained quickly (small \mathcal{M} and small τ). These situations are

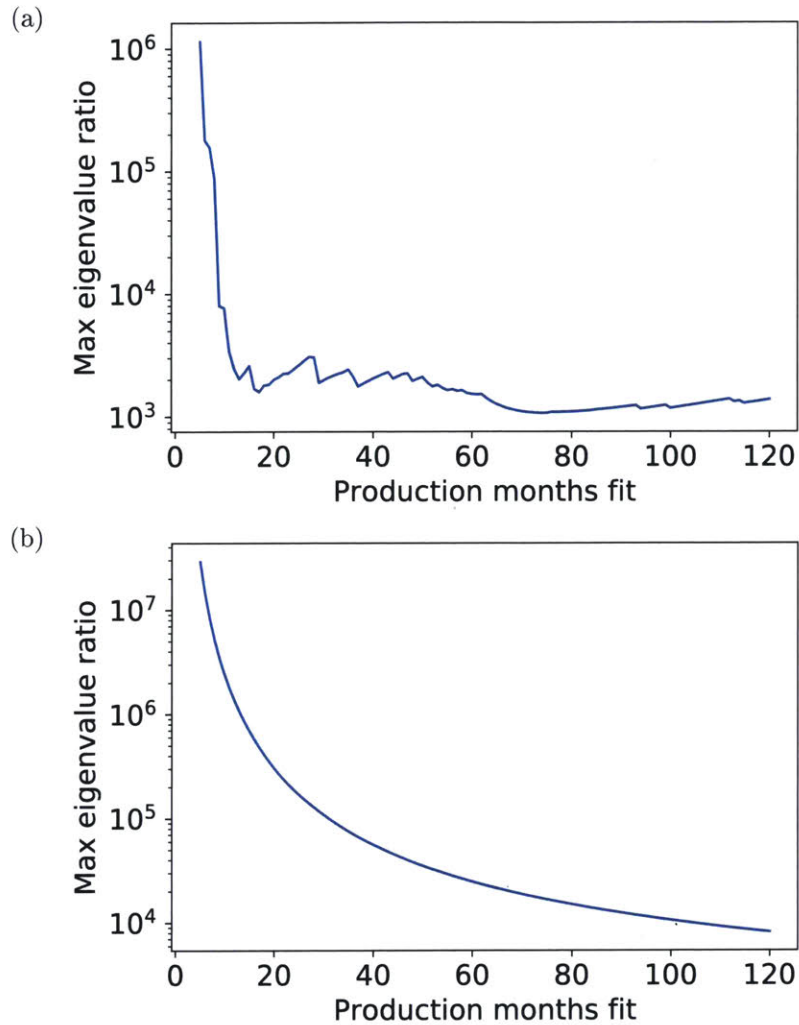


Figure 3-2: Sloppiness: The ratio of the largest to smallest eigenvalues of the Hessian for (a) The SCM and (b) The Arps curve.

indistinguishable based on only production data. Regularization reshapes the space toward more reliable parameter values.

Figure 3.2 and Fig. 3.2 show two slices of the loss function for the Arps curve applied to this same well. There is a similar extreme degree of parameter correlation, only in three dimensions now. Although these parameters are not physically interpretable they describe a similar kind of ambiguity which can be resolved using regularization to concentrate on a particular region in the parameter space.

When comparing the forecasting accuracy of these models, it is important to note that the SCM has two parameters while the modified Arps model has three (and the neural

network has nearly a thousand due to the dense connections between each layer). This influences how flexible the models are at fitting data, but here the focus is instead on each approach's reliability for making extrapolated predictions on out-of-sample test wells; the cross-validation approach helps avoid overfitting allowing for a fair comparison.

3.3 Discussion

For these representative DCA models, the step of introducing regularization matters more than model choice itself. This finding overshadows and somewhat trivializes the ongoing debate about which existing or new DCA model is best. The introduction of a DNN as a new model-neutral benchmark also provides a useful point of comparison. This approach has clear drawbacks from a practical forecasting perspective. It lacks interpretability or any physical meaning and requires abundant production data for training, making it less useful in new fields. However, the layering of nonlinear activation functions in the DNN creates effectively infinite functional flexibility while the noise in the representative training sets serves to adequately regularize the model and prevent overfitting [59]. As a result, it learns the best nonlinear mapping directly from the data, rather than presupposing an overly simplified model form. It can thus be used to establish the level of error attributable to underlying noise in the data rather than model mis-specification. With regularization, both DCA models approach on a similar level of accuracy as the DNN. Once the ill-posedness of the problem is addressed through regularization, the DCA models are able to predict production behavior reasonably well with far fewer parameters.

An interesting opportunity for further research is whether the neural network benchmark can be made more suitable for predictions in situations with less training data. One promising avenue for this is through transfer learning, in which a model is trained with a large dataset from one area and then only the final layers of the network are fine tuned with a smaller dataset in the area being predicted. This allows general knowledge from the larger dataset to be retained, while the more limited data specific to the prediction task is used to identify more granular patterns. An example of this is in image recognition, where there might be many training images of dogs but only a few of a specific breed. Transfer learning allows

a model to be first generally trained to identify dogs, and then additional layers used to identify the breed based on more granular features. In this context it could mean training a model in one extensively developed field to get a general understanding of production dynamics and then using limited samples in a new field to more rapidly identify patterns unique to the area.

There is an informative Bayesian perspective to regularization. Type-well curves are similar to a prior, since they are a static aggregation of field data and cannot be updated based on production dynamics of a well. DCA is a maximum likelihood estimate, completely uninformed by other wells in the field. Tikhonov regularization acts as a *max a posteriori* estimate, balancing these sources of information [94]. Bayesian modeling has in fact previously been proposed for quantifying uncertainty in DCA forecasts [32]. However, the findings here show that the sloppiness and ill-posedness of the likelihood makes this formulation critical for reliable point estimates too and the prior must be chosen with care. This suggests a promising direction for future work—hierarchical models that introduce physical information into DCA as conditional probabilities.

Ill-posed inverse problems are prevalent across many domains, including the biological and earth sciences [29, 59]. This analysis illustrates the importance of remaining vigilant to the challenges of forecasting with this class of problems. DCA (along with type-well curves) is the standard approach for unconventional oil and gas production forecasting today, behind large-scale resource outlooks and companies' reporting of reserves to investors [105, 93]. A lack of awareness of the ill-posed nature of DCA and the importance of regularization is unnecessarily introducing error into these projections and undermining the consequential decisions they inform.

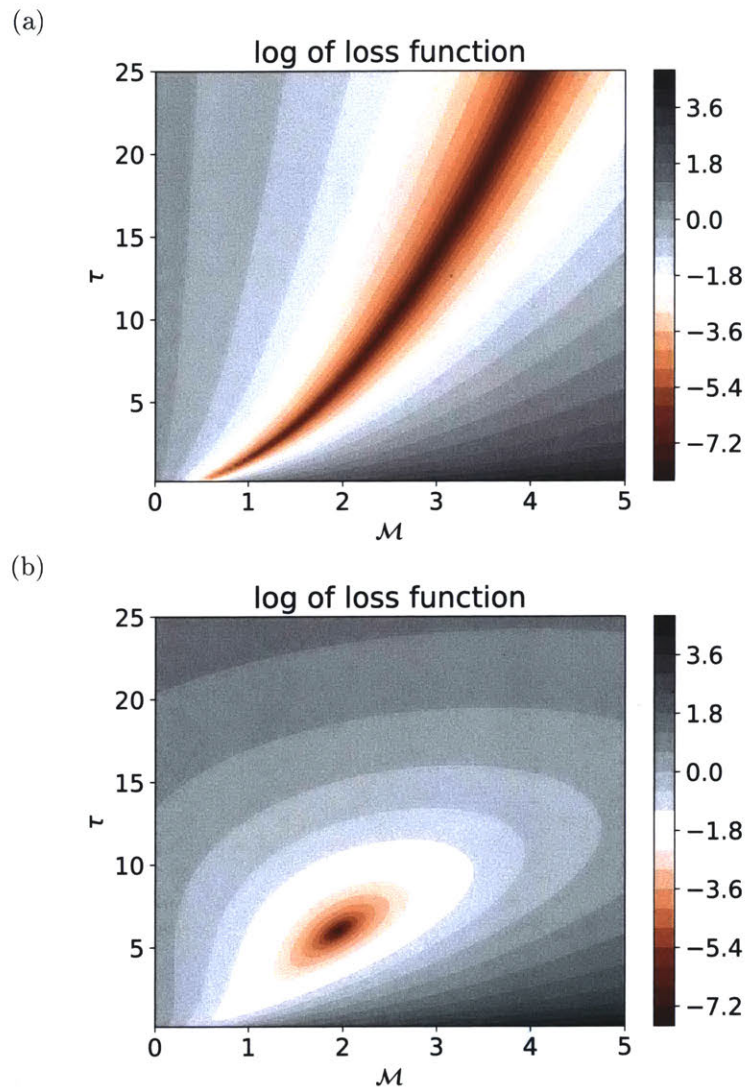


Figure 3-3: (a) Strong parameter correlation is apparent in the loss function associated with fitting the SCM to well production (shown here for the first 12 months of production in a well). (b) Regularization substantially reshapes the loss function by introducing prior information, reducing the ill-posedness of the problem.

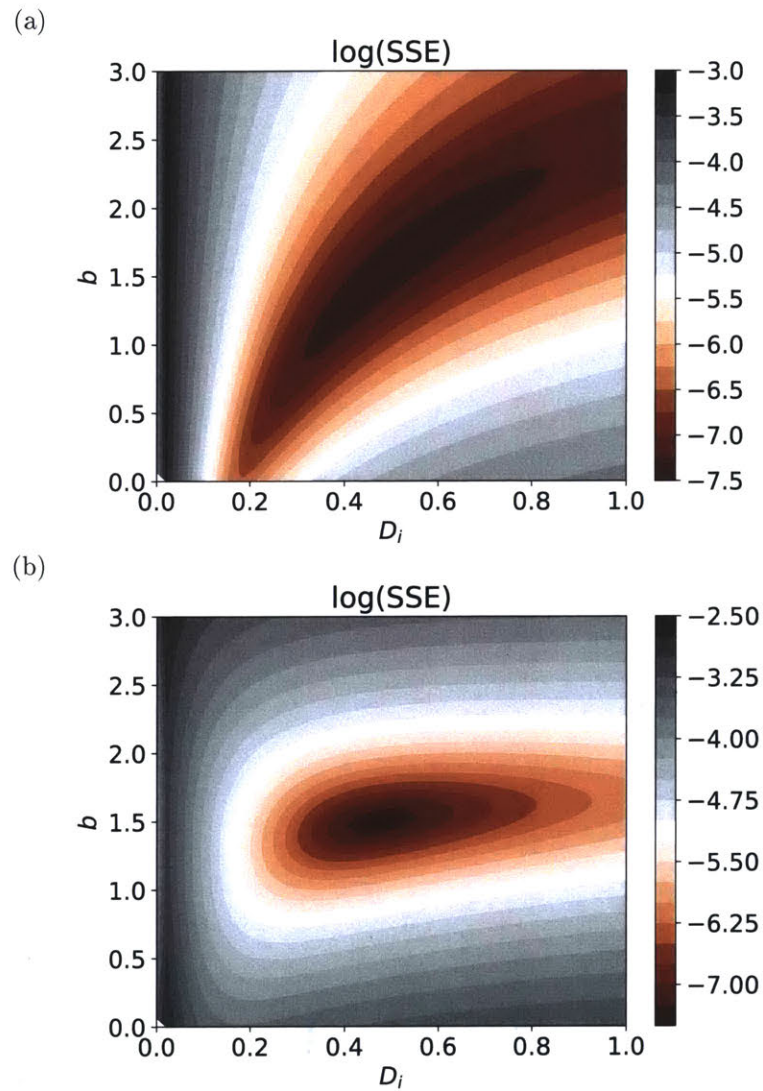


Figure 3-4: (a) Strong parameter correlation between b and D_i is apparent in the loss function associated with fitting the modified Arps curve to well production (shown here for the first 12 months of production in a well). (b) Regularization substantially reshapes the loss function by introducing prior information, reducing the ill-posedness of the problem.

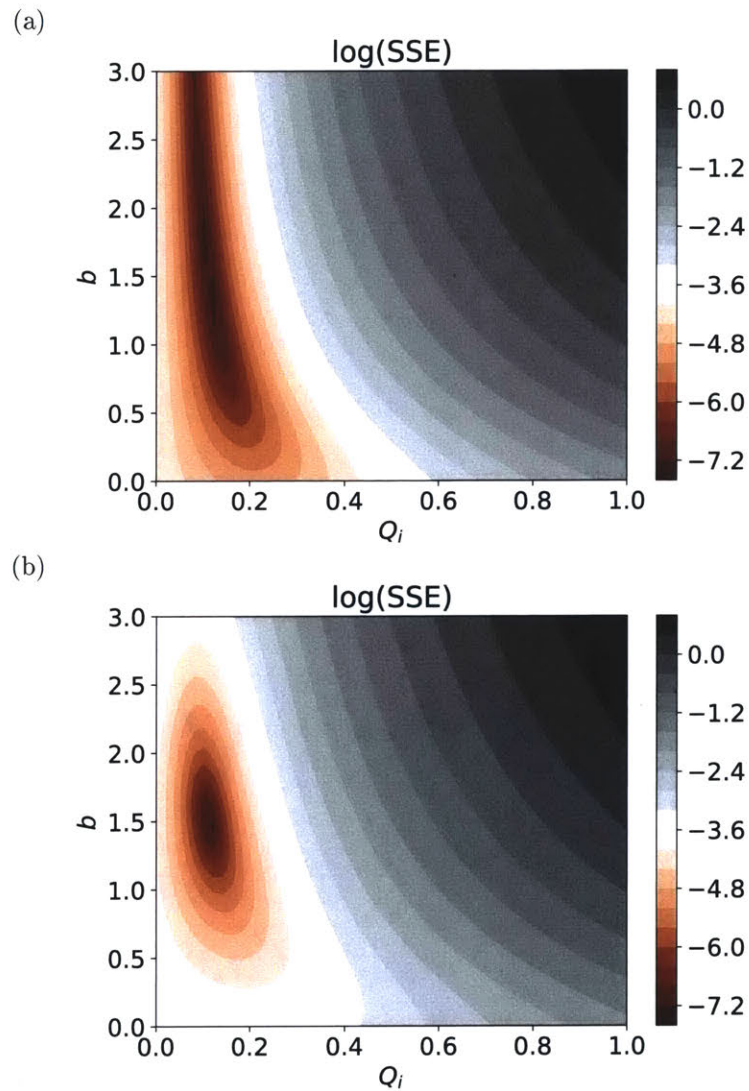


Figure 3-5: (a) Strong parameter correlation between b and Q_i is apparent in the loss function associated with fitting the modified Arps curve to well production (shown here for the first 12 months of production in a well). (b) Regularization substantially reshapes the loss function by introducing prior information, reducing the ill-posedness of the problem.

Chapter 4

A hierarchical Bayesian approach to incorporate physical information into mechanistic-statistical production forecasting

4.1 Introduction

Unconventional oil and gas development in North America has expanded rapidly over the past decade. Resource basins now contain thousands of producing wells and many more undrilled locations, but well productivity varies substantially, making it difficult to forecast future production levels [50, 61, 76]. Furthermore, evolving extraction techniques continue to reshape the dynamics of production considerably but current unconventional production forecasting approaches are unable to account for these physical differences across wells [68].

Unconventional production forecasting today is mostly by statistical rather than mechanistic modeling. Numerical reservoir simulations for unconventional wells are generally too costly and complex to be widely used, particularly given the challenges of adequately characterizing and modeling nano-scale flow in fractured reservoirs [110]. Instead, well production forecasts are typically based on a least-squares fit of a parameterized production curve model

(PCM) to the production rates observed in a well¹, as discussed in Chap. 3 [68, 95]. The estimated parameters are assumed to be representative of the conditions governing a well’s flow dynamics and are thus used to extrapolate production into the future. Meanwhile, production from other wells in the field is ignored despite potentially providing guidance on future production patterns. When limited production is available for fitting, as with a newer well, longer production histories from other wells in the field are usually aggregated into a type-well curve used as a proxy [105, 68]. These offset analogue wells may have significant physical differences compared to the well of interest but these factors are often ignored or used only as crude normalization factors for production, neglecting any impact on dynamics. A significant limitation of these forecasting approaches is the inability to rigorously balance production information from the well being forecast and other wells in the field while systematically controlling for physical differences based on other available data. Additionally, the inability to accurately portray uncertainty in these forecasts is a major hindrance to risk assessment and decision making today [68].

To address these shortcomings, this chapter introduces a spatial hierarchical Bayesian approach to better integrate and share information across all wells in a field when constructing production curve forecasts. Bayesian techniques have previously been used to quantify uncertainty in individual oil and gas well production forecasts by treating PCM parameters as probability distributions [23, 28, 32, 41, 114]. However, in all of these applications the prior distribution was either chosen to be uninformative or arbitrarily defined to provide a good match with historical production data, rather than using it to incorporate physical data. The main contribution of this chapter is the first Bayesian model for production that links wells through their shared dependence on underlying subsurface properties and mechanistic relationships. This kind of hierarchical approach has previously been used in other applications with noisy observations and known physical processes, including for paleoclimate temperature reconstruction from geothermal measurements by [16] and for fishery biomass dynamics by [72]. It has also been recognized as a particularly advantageous framework for spatial environmental and geophysical problems plagued by uncertainty [9, 111].

¹In Chap. 3, this was referred to as decline curve analysis (DCA), as the approach is typically called in industry. In this chapter, it is more appropriate to refer to the PCM directly because the terminology “DCA” is so strongly associated with least squares fitting.

As physical differences between unconventional wells have become more pronounced, it is increasingly important to understand how these differences impact long-term productivity in order to create reliable forecasts and inform decisions in well designs. Fortunately, this mechanistic-statistical hierarchical approach offers the potential to infer these relationships across an entire population of wells, advancing our empirical understanding of well production behavior and as a result enabling better forecasts for new, physically different wells.

As a result of the extremely low permeability of unconventional oil and gas reservoirs, production dynamics are dominated by the properties of the stimulated reservoir volume (SRV)—the region where hydraulic fracturing has opened conduits in the rock to allow oil and gas to flow to the well [49, 109]. Pursuit of a larger and enhanced SRV in wells has driven recent well designs toward longer horizontal producing segments and increased volumes of water used in hydraulic fracturing [103]. It is because the SRV properties are so uncertain and cannot be directly observed that mechanistic modeling of production based only on prior knowledge is infeasible [84]. However, the mechanistic-statistical approach infers these properties by combining production data with abundant information about technical design specifications of wells and estimated geological properties which influence this SRV.

To establish a mechanistic relationship between production behavior and the SRV, the hierarchical model leverages the Scaling Curve Model (SCM), recently developed by [84] based on the non-dimensionalized solution to a one-dimensional flow problem with planar fractures as illustrated in Fig. 4-1. This model was introduced in Chap. 3. The cumulative production $Q(t)$ at time t is parameterized using the nonlinear relationship

$$Q(t) = \mathcal{M}F(t/\tau), \quad (4.1)$$

where $F(\cdot)$ is the forward model or, more specifically, the tabulated solution to the partial differential equation describing the flow behavior. In this model, wells are assumed to drain gas only from the SRV region between planar fractures of height H , half-length L , and a distance of $2d$ apart. With N_s stages in a horizontal well, the total SRV is $4(N_s + 1)LHd$ [84]. The producible mass of gas \mathcal{M} depends on the SRV as well as the porosity ϕ , gas saturation

S_g , and density ρ as in

$$\mathcal{M} = 4(N_s + 1)LHd\phi S_g\rho. \quad (4.2)$$

Assuming linear flow, the pressure interference time τ between fractures is

$$\tau = \frac{d^2\phi S_g\mu_g c_g}{k_{SRV}}, \quad (4.3)$$

where gas viscosity is μ_g , gas compressibility is c_g , and effective enhanced reservoir permeability within the SRV is k_{SRV} . Production declines as a square root of time until this interference time is reached and then exponentially thereafter. This simplified physical model neglects desorption and nonlinear flow behavior but captures the dominant mechanisms of production and provides a good empirical fit to production data for thousands of wells in a shale gas field, as shown by [84].

The SCM provides a direct theoretical connection between PCM parameters estimated from production data and the properties and geometry of the SRV. However, the inherent uncertainty about important subsurface properties in Eq. 4.2 and Eq. 4.3 has thus far made it difficult to use the SCM to develop production forecasts early in the life of a well. Without a reliable estimate for the SRV, it is impossible to reach a unique estimate of the parameters \mathcal{M} and τ from limited production data. High initial production rates could indicate a large SRV, but could also plausibly be a result of a much smaller SRV being drained quickly, precipitating a more rapid drop in production over time [84]. This ambiguity makes forecasting

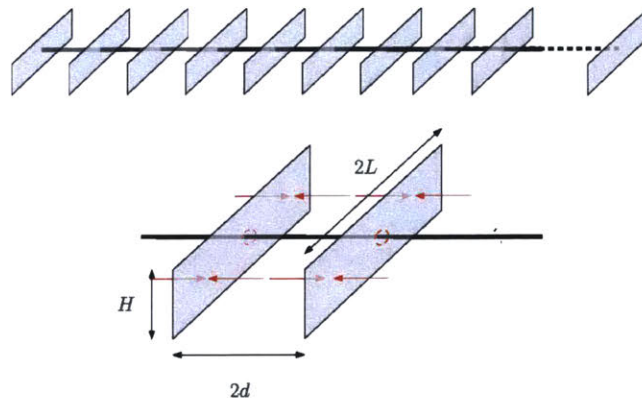


Figure 4-1: Horizontal well and planar fracture geometry assumed in the simplified one-dimensional flow model. Figure is from [84]

with the scaling curve an underidentified or ill-posed inverse problem which requires further information from the field to resolve, something the hierarchical approach is ideally suited for [30, 67].

The SCM was derived specifically for the physical conditions of the Barnett shale gas field and is not directly applicable to other fields. The hierarchical approach is flexible though, and can be adapted to work with other PCMs. A slightly more flexible three-parameter empirical model called the Logistic Growth Model (LGM) captures similar temporal behavior while maintaining the direct connection to the SRV. The LGM has roots in modeling population growth [107] but can be more generally applied to dynamic systems with saturating behavior [99]. A variant was used by [42] to describe production from entire fields and was adapted for individual unconventional oil and gas wells by [18] as

$$Q(t) = \frac{\mathcal{K}t^\eta}{\nu + t^\eta}. \quad (4.4)$$

The cumulative production $Q(t)$, increases over time t toward a carrying capacity, \mathcal{K} , at a rate jointly determined by η and ν . Although the LGM is not derived from physics like the SCM, there is a direct analogy between the role of the \mathcal{M} and \mathcal{K} parameters in defining total potential production from a well.

In fact, closer comparison of the SCM and LGM shows that they bear remarkable similarities despite very different origins in physical theory and population dynamics. The LGM, with its three parameters, can easily match the behavior of the SCM, as shown in Fig 4-2 where the LGM has been fit using nonlinear least squares to a realization of the SCM with typical parameters. Carrying this out across a range of parameters in the SCM and examining the correspondence between parameters in each model reveals the underlying relationship between models, shown in Fig. 4.1. For a fixed value of τ , there is a linear mapping of \mathcal{M} to \mathcal{K} , where the slope is influenced by τ . The parameters η and ν have no connection to \mathcal{M} but are related to τ through a consistent nonlinear relationship. As Fig. 4.1(b) shows, an approximation of this nonlinear relationship is $\eta^{1/\nu} \propto \tau^2$ with some deviation at lower values. Understanding the principles that lead a function for population dynamics to so strongly resemble a partial differential equation governing fractured gas flow behavior is

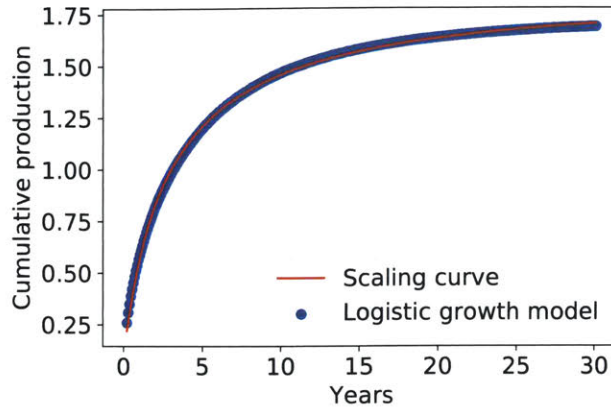


Figure 4-2: LGM fit using nonlinear least squares to realization of SCM with typical parameters

beyond the scope of this thesis. For the purposes here, it is enough to recognize that the more flexible LGM provides an adequate representation for mechanistic flow behavior and is worth considering as an alternative model alongside the SCM.

This chapter provides a general methodology in Sect. 4.2 for mechanistic-statistical production forecasting using hierarchical Bayesian formulations of the SCM and LGM. As with [72], a metropolis-Hastings (MH) within Gibbs sampler is employed to efficiently approximate the posterior with Markov Chain Monte Carlo (MCMC) samples. This sampling scheme is described in Sect. 4.2.4 along with steps to take advantage of additional structure and correlation in the likelihood function when generating MH samples. The model is able to incorporate additional sources of geological and engineering data for applications in both a shale gas and tight oil field, as described in Sect. 4.3.1. In Sect. 4.3.2, the interpretability of the model is shown using marginal distributions of parameters. Additionally, the accuracy of predictions in two-fold cross-validation is compared to a traditional nonlinear least-squares forecast and a field averaged type-well curve prediction. By striking a balance between data from an individual well and its neighbors in a partially-pooled manner, accuracy is substantially enhanced and the range of uncertainty is narrowed. The error introduced by approximating the hierarchical model with an empirical Bayesian approach is shown to be small but increases when the test set of wells is substantially different from the other wells in the population. Finally, Sect. 4.4 includes some concluding remarks and promising areas of future work.

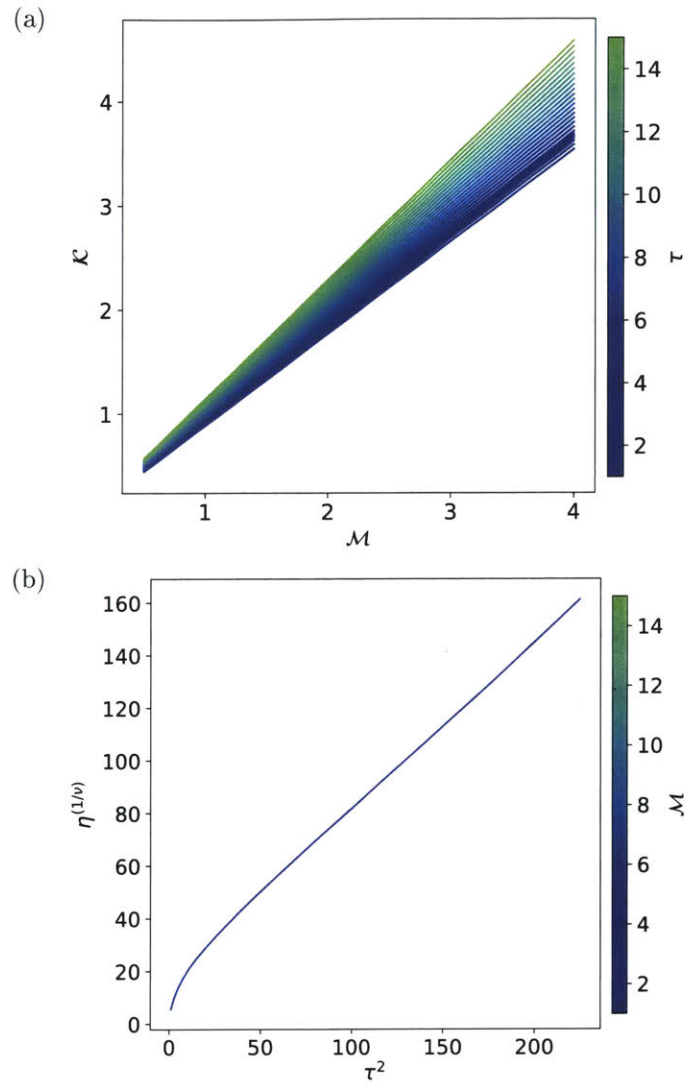


Figure 4-3: Parameter mappings between the SCM and LGM found by fitting the LGM to realizations of the SCM for different parameter realizations. (a) Mapping of \mathcal{M} to \mathcal{K} for different values of τ . (b) Mapping of τ^2 to $\eta^{1/\nu}$ for different values of \mathcal{M} (lines are on top of each other because \mathcal{M} has no impact).

4.2 Methods

4.2.1 Least-squares and type-well curves

The standard approach to forecasting with PCMs, including the SCM in Eq. 4.1 and the LGM in Eq. 4.4, is using a nonlinear least squares algorithm to estimate the parameters that minimize the total squared mismatch between the PCM and production data, typically observed on a monthly basis [18, 84, 95]. This least-squares minimization has been implemented using a standard Levenberg-Marquardt algorithm in the Python package `lmfit`, following the methodology of [84]. A standard type-well curve approach has also been implemented for comparison. For this, production time series from offset training wells are aggregated and the mean is used as a prediction for new wells.

4.2.2 Bayesian Formulation for Production Curve Models

For a particular well i , the least-squares approach is equivalent to assuming Gaussian noise in the production time series (t_{im}, Q_{im}) for each monthly observation $m = 1 \dots M_i$ where M_i is the total months of production for that well. A least squares estimate then seeks the PCM parameters—such as $\{\mathcal{M}_i, \tau_i\}$ when the SCM is used—that maximize the likelihood distribution for this well’s production data. As production is lower bounded by zero, this assumption is slightly more appropriate when the PCM is first log-transformed, leading to a likelihood function for a well i of

$$\begin{aligned}
 p(\{Q_{im}\}_{m=1}^{M_i} | \mathcal{M}_i, \tau_i, \{t_{im}\}_{m=1}^{M_i}) \\
 = \prod_{m=1}^{M_i} \mathcal{N}(\log(Q_{im}); \log(\mathcal{M}_i) + \log(F(t_{im}/\tau_i)), \sigma_Q^2). \quad (4.5)
 \end{aligned}$$

for the SCM and

$$\begin{aligned}
 p(\{Q_{im}\}_{m=1}^{M_i} | \mathcal{K}_i, \eta_i, \nu_i, \{t_{im}\}_{m=1}^{M_i}) \\
 = \prod_{m=1}^{M_i} \mathcal{N}\left(\log(Q_{im}); \log(\mathcal{K}_i) - \log\left(\frac{\eta_i}{t_{im}^{\nu_i}} + 1\right), \sigma_Q^2\right) \quad (4.6)
 \end{aligned}$$

for the LGM. The notation $\{Q_{im}\}_{m=1}^{M_i}$ refers to the set of cumulative production values $(Q_{i1}, Q_{i2}, \dots, Q_{iM_i})$ and likewise for $\{t_{im}\}_{m=1}^{M_i}$. To be more concise going forward, these time series will be written as $(\mathbf{t}_i, \mathbf{Q}_i)$. Additionally, $\mathcal{N}(x; \mu, \sigma^2)$ is used to denote the probability density of a value x in a Gaussian distribution with mean μ and variance σ^2 .

As the parameters for both the SCM and LGM must be positive values, it can be convenient to work with log-transformed parameters so $\boldsymbol{\theta}_i$ is used to denote either $\{\log(\mathcal{M}_i), \log(\tau_i)\}$ for the SCM or $\{\log(\mathcal{K}_i), \log(\eta_i), \log(\nu_i)\}$ for the LGM. The Bayesian approach to inferring a well's PCM parameters $\boldsymbol{\theta}_i$ is to combine the likelihood in Eq. 4.5 or 4.6 with a prior distribution for $\boldsymbol{\theta}_i$ according to Bayes' rule. This posterior distribution π can be written as

$$\pi(\boldsymbol{\theta}_i | \mathbf{t}_i, \mathbf{Q}_i) = \frac{p(\mathbf{Q}_i | \boldsymbol{\theta}_i, \mathbf{t}_i)p(\boldsymbol{\theta}_i)}{p(\mathbf{t}_i, \mathbf{Q}_i)} \quad (4.7)$$

$$\propto p(\mathbf{Q}_i | \boldsymbol{\theta}_i, \mathbf{t}_i)p(\boldsymbol{\theta}_i) \quad (4.8)$$

where it is often easier to work with the unnormalized proportionality in Eq. 4.8. The resulting posterior distribution $p(\boldsymbol{\theta}_i | \mathbf{t}_i, \mathbf{Q}_i)$ defines the probability of any combination of PCM parameters given the production data and prior distribution. As the posterior is often intractable, it is standard to use Markov Chain Monte Carlo (MCMC) approaches to generate samples approximating it. This is discussed in detail in Sect. 4.2.4.

Choosing the prior distribution is an important and nuanced part of Bayesian modeling and is useful for introducing other information and beliefs into the probability model. When there are many model parameters that are potentially related or structurally linked, this dependence can be captured using hierarchical prior relationships based on conditional probability [30]. For instance, each individual well i and its associated PCM parameters $\boldsymbol{\theta}_i$ can be viewed as being drawn from a common distribution for the entire population of W producing wells in the unconventional basin or field. Given the challenge of working with high-dimensional Bayesian models, a less fully-Bayesian approach—typically referred to as an empirical Bayesian model—is often used to approximate the hierarchical prior based on an analysis of data from other individuals in the population. The empirical Bayesian approach is implemented here by assuming independent Gaussian priors on each dimension of $\boldsymbol{\theta}$ with the means and variances based on the empirical distribution of nonlinear least squares

estimates of θ_i for each well (Sect. 4.2.1). As shown in Sect. 4.3.2, this approximation may be adequate for some situations but it lacks the generalizability and insight offered by the more fully Bayesian approach developed in the following section.

4.2.3 Hierarchical formulation with Gaussian process

A hierarchical model can be constructed for the W producing wells in a field by recognizing correlations and dependence between the PCM parameters for each well and other physical properties. This is developed here for the SCM parameter \mathcal{M} but the same formulation applies to \mathcal{K} in the LGM since it similarly determines the total producible resource for a well.

The producible mass of gas for a well \mathcal{M}_i is determined by both spatially varying geological properties in the field and design specifications of the well which influence the SRV geometry in Eq. 4.2. For known geological properties and well design parameters, a linear relationship to $\log(\mathcal{M}_i)$ can be assumed, as in

$$\log(\mathcal{M}_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad (4.9)$$

where ϵ_i is distributed according to a zero-mean Gaussian distribution, \mathbf{x}_i^T is a $(P + 1)$ row vector of the p covariates and a 1 for the intercept, and $\boldsymbol{\beta}$ is the corresponding $(P + 1)$ column vector of coefficients. Many important underlying geological properties are unknown but similar for wells that are near each other. This leads to spatial autocorrelation of ϵ_i across wells, which can significantly bias estimates of $\boldsymbol{\beta}$ if ignored, as shown in Chap. 2. However, utilizing the information contained in this spatial autocorrelation leads to the powerful predictive technique of Gaussian process regression, or kriging as it is called in spatial statistics [22, 74]. This spatial autocorrelation can be described using a covariance

function C , as in

$$C(\mathbf{s}_i, \mathbf{s}_{i'}) = \text{Cov}[\log(\mathcal{M}_i), \log(\mathcal{M}_{i'})] \quad (4.10)$$

$$= \text{E}[(\log(\mathcal{M}_i) - \mathbf{x}_i^T \boldsymbol{\beta})(\log(\mathcal{M}_{i'}) - \mathbf{x}_{i'}^T \boldsymbol{\beta})] \quad (4.11)$$

$$= \tilde{C}(\mathbf{s}_i - \mathbf{s}_{i'}), \quad (4.12)$$

where $\mathbf{s}_i \in \mathbb{R}^2$ is each well's coordinate location [67]. Following [74], the slightly rough exponential covariance function

$$\tilde{C}(\mathbf{s}_i - \mathbf{s}_{i'}) = \gamma_1 \exp\left(-\frac{h}{\gamma_2}\right) \quad (4.13)$$

provides an appropriate model for describing the nature of spatial autocorrelation in unconventional well productivity. This treats the covariance as isotropic, depending only on the distance h between \mathbf{s}_i and $\mathbf{s}_{i'}$. The γ parameters determine the variance and length scale of the covariance function and [74] described how they can be estimated using a semivariogram fitting procedure. Here they are taken as known since they are hyperparameters of the prior that can be reasonably estimated from the nonlinear least squares estimates of $\log(\mathcal{M}_i)$ [67].

The prior for $\log(\mathcal{M}_i)$ can thus be distributed according to the hierarchical prior

$$\log(\mathcal{M}_i) \sim \mathcal{GP}(\mathbf{x}_i^T \boldsymbol{\beta}, C(\mathbf{s}_i, \mathbf{s}_{i'}) + \sigma_{\mathcal{M}}^2), \quad (4.14)$$

where the Gaussian process \mathcal{GP} is parameterized by the mean trend $\mathbf{x}_i^T \boldsymbol{\beta}$ from Eq. 4.9 and covariance kernel $C(\mathbf{s}_i, \mathbf{s}_{i'})$. The parameter $\sigma_{\mathcal{M}}^2$ defines the level of shrinkage, or how strongly values of $\log(\mathcal{M}_i)$ should be pulled toward the mean of the Gaussian process.

Treating the value of the Gaussian process at each well in the field as a random variable is computationally burdensome due to the potentially large number of wells in the field. Instead, a truncated Karhunen-Loève (KL) expansion can be used to approximate the Gaussian process [67], as

$$\mathcal{GP}(\mathbf{x}_i^T \boldsymbol{\beta}, C(\mathbf{s}_i, \mathbf{s}_{i'})) \approx \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{k=1}^K \omega_k \varphi_k(\mathbf{s}_i) \sqrt{\lambda_k}, \quad (4.15)$$

where the stochasticity of the Gaussian process is obtained using independent standard Gaussian random variables, ω_k , as weights for each mode $k = 1 \dots K$. The eigenvalues and eigenfunctions to the covariance Kernel, λ_k and $\varphi_k(\mathbf{s}_i)$, solve

$$\int_{\mathbb{D}} C(\mathbf{s}_i, \mathbf{s}_{i'}) \varphi_k(\mathbf{s}_{i'}) d\mathbf{s}_{i'} = \lambda_k \varphi_k(\mathbf{s}_i), \quad (4.16)$$

where \mathbb{D} is the domain containing \mathbf{s}_i . By Mercer's theorem, they also form the orthogonal decomposition of the covariance kernel:

$$C(\mathbf{s}_i, \mathbf{s}_{i'}) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(\mathbf{s}_i) \varphi_n(\mathbf{s}_{i'}). \quad (4.17)$$

The first N eigenvalues and eigenfunctions of Eq. 4.16 can be found numerically using the Nystrom method [87]. This has been implemented for KL expansions of covariance kernels in the MUQ2 open-source software, which is used here (web: muq.mit.edu). Analysis of the eigenvalues decay indicates the error introduced by truncation of higher order terms, and $N = 40$ modes was deemed to be a suitable approximation for the covariance kernels used here.

It is convenient to recognize that when the covariance kernel and number of modes are fixed, the values of $\varphi_k(\mathbf{s}_i) \sqrt{\lambda_k}$ can be pre-computed for each location \mathbf{s}_i and the truncated KL expansion in Eq. 4.15 can be restated as

$$\mathcal{GP}(\mathbf{x}_i^T \boldsymbol{\beta}, C(\mathbf{s}_i, \mathbf{s}_{i'})) \approx \boldsymbol{\zeta}^T \boldsymbol{\Psi}_i = \sum_{j=1}^J \zeta_j \Psi_{ij}, \quad (4.18)$$

where $J = (P+1+N)$ and $\boldsymbol{\zeta}$ is the J -length column vector combining the $P+1$ dimensional $\boldsymbol{\beta}$ with the N dimensional $\boldsymbol{\omega}$, as in $\boldsymbol{\zeta}^T = [\boldsymbol{\beta}^T, \boldsymbol{\omega}^T]$. The covariates \mathbf{x}_i are also combined with the pre-computed KL modes in $\varphi_k(\mathbf{s}_i) \sqrt{\lambda_k}$ to form the $J \times 1$ column vector $\boldsymbol{\Psi}_i$, as in $\boldsymbol{\Psi}_i^T = [\mathbf{x}_i^T, (\varphi_k(\mathbf{s}_i) \sqrt{\lambda_k})^T]$.

The prior on each of the coefficients ζ_j is chosen to be Gaussian

$$p(\zeta_j) = \mathcal{N}(\zeta_j; \mu_{\zeta_j}, \sigma_{\zeta_j}^2). \quad (4.19)$$

It is assumed that μ_{ζ_j} is zero with $\sigma_{\zeta_j}^2 = 0.1$ for the coefficients corresponding to KL modes. A more informative prior can be used for the dimensions of ζ corresponding to β by using the least squares parameter estimates in Sect. 4.2.1 to estimate these μ_{ζ_j} . Following the discussion in Sect. 4.2.2, empirical Gaussian priors are still used for the other log-transformed parameters in the PCMs, such as for τ_i in the SCM:

$$p(\log(\tau_i)) = \mathcal{N}(\log(\tau_i); \mu_\tau, \sigma_\tau^2). \quad (4.20)$$

The joint posterior over all W wells is

$$\begin{aligned} & p(\{\log(\mathcal{M}_i)\}_{i=1}^W, \{\log(\tau_i)\}_{i=1}^W, \{\zeta_j\}_{j=1}^J | \{\mathbf{t}_i\}_{i=1}^W, \{\mathbf{Q}_i\}_{i=1}^W, \{\mathbf{s}_i\}_{i=1}^W, \{\mathbf{x}_i\}_{i=1}^W) \\ & \quad \propto p(\{\mathbf{Q}_i\}_{i=1}^W | \{\log(\tau_i)\}_{i=1}^W, \{\log(\mathcal{M}_i)\}_{i=1}^W, \{\mathbf{t}_i\}_{i=1}^W) \\ & \quad p(\{\log(\mathcal{M}_i)\}_{i=1}^W | \{\zeta_j\}_{j=1}^J, \{\mathbf{s}_i\}_{i=1}^W, \{\mathbf{x}_i\}_{i=1}^W) p(\{\zeta_j\}_{j=1}^J) p(\{\log(\tau_i)\}_{i=1}^W) \\ & \quad \propto \prod_{i=1}^W \left(p(\mathbf{Q}_i | \log(\tau_i), \log(\mathcal{M}_i), \mathbf{t}_i) \right) \\ & \quad p(\{\log(\mathcal{M}_i)\}_{i=1}^W | \{\zeta_j\}_{j=1}^J, \{\mathbf{s}_i\}_{i=1}^W, \{\mathbf{x}_i\}_{i=1}^W) p(\{\zeta_j\}_{j=1}^J) p(\{\log(\tau_i)\}_{i=1}^W). \end{aligned} \quad (4.21)$$

for the SCM and is the same for the LGM except $\log(\mathcal{M}_i)$ is replaced by $\log(\mathcal{K}_i)$ and $\log(\tau_i)$ is replaced by the parameters $\log(\eta_i)$ and $\log(\nu_i)$.

4.2.4 Sampling using Metropolis within Gibbs

The hierarchical model's joint posterior in Eq. 4.21 is intractable, as is typical of complex probabilistic models. A practical approach to inference with such models is to employ a Markov chain Monte Carlo (MCMC) sampler to generate samples asymptotically approximating the posterior distribution. The Metropolis-Hastings (MH) algorithm [38] is a general method for MCMC. For each step $n = 1 \dots N$, a proposal z' is drawn from a proposal distribution $q(z' | z^{(n)})$ where $z^{(n)}$ is the current state in the chain. The acceptance ratio is calculated using the posterior distribution π as

$$\alpha(z^{(n)}, z') = \min \left\{ 1, \frac{\pi(z') q(z^{(n)} | z')}{\pi(z^{(n)}) q(z' | z^{(n)})} \right\} \quad (4.22)$$

and the chain accepts the proposal and advances to $z^{(n+1)} = z'$ with probability $\alpha(z^{(n)}, z')$ or spends longer in the current state, as in $z^{(n+1)} = z^{(n)}$, with probability $1 - \alpha(z^{(n)}, z')$. With a sufficient number of steps N , the stationary distribution of the chain converges to π regardless of the initial value at $z^{(0)}$.

Even with the dimensionality reduction enabled by approximating the Gaussian process with a truncated Karhunen-Loève expansion, the number of random variables in the hierarchical model remains very large. For example, with $N = 40$ modes of the KL expansion retained, $P = 4$ regression parameters, and $N = 1000$ wells, there are 3045 random variables with the LGM or 2045 for the SCM. This makes generic Metropolis-Hastings sampling schemes prohibitive and it is essential to exploit the structure of the problem to achieve more efficient sampling. One approach is using Gibbs sampling for parts of the model with a structure that allows the full conditional distribution to be derived and sampled from directly. This is equivalent to using the full-conditional distribution as the proposal distribution in MH leading to an acceptance rate of 1 in Eq. 4.22 [30]. For variables where the full-conditional cannot be derived, MH can be used, leading to the Metropolis within Gibbs algorithm.

In this problem, Gibbs sampling is used for ζ_j and $\log(\mathcal{K}_i)$. The full-conditional for each ζ_j in the SCM is found by retaining only the parts of Eq. 4.21 with conditional dependence on ζ_j , as in

$$\begin{aligned}
p\left(\zeta_j \mid \{\mathbf{t}_i\}_{i=1}^W, \{\mathbf{Q}_i\}_{i=1}^W, \{\mathbf{\Psi}_i\}_{i=1}^W, \{\boldsymbol{\theta}_i\}_{i=1}^W, \{\zeta_\ell\}_{\ell \geq 1, \ell \neq j}^J\right) \\
\propto p\left(\{\log(\mathcal{M}_i)\}_{i=1}^W \mid \{\zeta_\ell\}_{\ell=1}^J, \{\mathbf{\Psi}_i\}_{i=1}^W\right) p(\zeta_j) \\
= \prod_{i=1}^W \mathcal{N}\left(\log(\mathcal{M}_i); \boldsymbol{\zeta}^T \mathbf{\Psi}_i, \sigma_{\mathcal{M}}^2\right) \mathcal{N}\left(\zeta_j; \mu_{\zeta_j}, \sigma_{\zeta_j}^2\right) \quad (4.23)
\end{aligned}$$

Taking the logarithm of Eq. 4.23, again keeping only the terms involving ζ_j , and collecting

the ζ_j and ζ_j^2 terms separately gives

$$\begin{aligned}
& \log [p(\{\log(\mathcal{M}_i)\}_{i=1}^W | \{\zeta_\ell\}_{\ell=1}^J, \{\Psi_i\}_{i=1}^W) p(\zeta_j)] \\
& \propto -\frac{1}{2\sigma_{\mathcal{M}}^2} \sum_{i=1}^W [\log(\mathcal{M}_i) - \zeta^T \Psi_i]^2 - \frac{(\zeta_j - \mu_{\zeta_j})^2}{2\sigma_{\zeta_j}^2} \\
& \propto \frac{1}{\sigma_{\mathcal{M}}^2} \sum_{i=1}^W \left[\Psi_{ij} \log(\mathcal{M}_i) - \sum_{\substack{1 \leq \ell \leq J \\ \ell \neq j}} \zeta_\ell \Psi_{i\ell} \Psi_{ij} \right] \zeta_j - \frac{1}{2} \left[\frac{1}{\sigma_{\zeta_j}^2} + \frac{1}{\sigma_{\mathcal{M}}^2} \sum_{i=1}^W \Psi_{ij}^2 \right] \zeta_j^2. \quad (4.24)
\end{aligned}$$

Because the log-conditional posterior dependence on ζ_j matches the quadratic form of a Gaussian distribution we can identify the Gibbs updated distribution of ζ'_j as Gaussian with precision $\tau_{\zeta'_j}$ as

$$\frac{1}{\sigma_{\zeta'_j}^2} = \tau_{\zeta'_j} = \frac{1}{\sigma_{\zeta_j}^2} + \frac{1}{\sigma_{\mathcal{M}}^2} \sum_{i=1}^W \Psi_{ij}^2 \quad (4.25)$$

and mean $\mu_{\zeta'_j}$ defined by

$$\mu_{\zeta'_j} = \frac{1/\sigma_{\mathcal{M}}^2 \sum_{i=1}^W \left[\Psi_{ij} \log(\mathcal{M}_i) - \sum_{\ell \geq 1, \ell \neq j}^J \zeta_\ell \Psi_{i\ell} \Psi_{ij} \right] + \mu_{\zeta_j} / \sigma_{\zeta_j}^2}{\tau_{\zeta'_j}}. \quad (4.26)$$

The same derivation applies to the LGM but with $\log(\mathcal{K}_i)$ substituted for $\log(\mathcal{M}_i)$.

The full conditional for Gibbs updating each $\log(\mathcal{K}_i)$ is

$$\begin{aligned}
& p(\log(\mathcal{K}_i) | \{\mathbf{t}_i\}_{i=1}^W, \{\mathbf{Q}_i\}_{i=1}^W, \{\Psi_i\}_{i=1}^W, \{\log(\eta_i)\}_{i=1}^W, \{\log(\nu_i)\}_{i=1}^W, \{\zeta_j\}_{j=1}^J) \\
& \propto p(\mathbf{Q}_i | \log(\eta_i), \log(\nu_i), \log(\mathcal{K}_i), \mathbf{t}_{im}) p(\log(\mathcal{K}_i) | \{\zeta_j\}_{j=1}^J, \{\Psi_i\}_{i=1}^W). \quad (4.27)
\end{aligned}$$

Note that $\log(\mathcal{K}_i)$ is conditionally independent of the LGM parameters for all other wells in the field: $\log(\mathcal{K}_{\sim i})$, $\log(\eta_{\sim i})$, and $\log(\nu_{\sim i})$. The log-conditional posterior distribution is

then

$$\begin{aligned}
& \log [p(\mathbf{Q}_i | \log(\eta_i), \log(\nu_i), \log(\mathcal{K}_i), \mathbf{t}_i) p(\log(\mathcal{K}_i) | \{\zeta_j\}_{j=1}^J, \{\Psi_i\}_{i=1}^W)] \\
& \propto \frac{-1}{2\sigma_Q^2} \sum_{m=1}^{M_i} \left[\log(Q_{im}) - \log(\mathcal{K}_i) + \log\left(\frac{\eta_i}{t_{im}^{\nu_i}} + 1\right) \right]^2 - \frac{1}{2\sigma_{\mathcal{K}}^2} [\log(\mathcal{K}_i) - \boldsymbol{\zeta}^T \boldsymbol{\Psi}_i]^2 \\
& \propto \left(\frac{1}{\sigma_Q^2} \sum_{m=1}^{M_i} \left[\log\left(\frac{\eta_i}{t_{im}^{\nu_i}} + 1\right) + \log(Q_{im}) \right] + \frac{\boldsymbol{\zeta}^T \boldsymbol{\Psi}_i}{\sigma_{\mathcal{K}}^2} \right) \log(\mathcal{K}_i) \\
& \quad - \frac{1}{2} \left(\frac{M_i}{\sigma_Q^2} + \frac{1}{\sigma_{\mathcal{K}}^2} \right) \log(\mathcal{K}_i)^2. \quad (4.28)
\end{aligned}$$

From this, the precision and mean for the Gibbs update of $\log(\mathcal{K}'_i)$ can be analytically determined as

$$\frac{1}{\sigma_{\mathcal{K}'_i}^2} = \tau_{\mathcal{K}'_i} = \frac{M_i}{\sigma_Q^2} + \frac{1}{\sigma_{\mathcal{K}}^2} \quad (4.29)$$

and

$$\mu_{\mathcal{K}'_i} = \frac{1/\sigma_Q^2 \sum_{m=1}^{M_i} [\log(\eta_i/t_{im}^{\nu_i} + 1) + \log(Q_{im})] + [\boldsymbol{\zeta}^T \boldsymbol{\Psi}_i]/\sigma_{\mathcal{K}}^2}{\tau_{\mathcal{K}'_i}}. \quad (4.30)$$

To aid with the efficiency of sampling from the remaining variables, two other techniques are demonstrated which utilize other structure in the model. For the SCM, it is highly efficient to take advantage of the high degree of parameter correlation between $\log(\mathcal{M}_i)$ and $\log(\boldsymbol{\tau}_i)$ using a standard adaptive Metropolis-Hastings (AMH) algorithm [36]. This AMH algorithm was similarly employed for inference on individual well PCMs by [114]. The pair of parameters in $\boldsymbol{\theta}_i$ for each well i are sampled as a block using a MH bivariate Gaussian proposal centered at the current state in the chain $\boldsymbol{\theta}_i^{(n)}$. For the first n_0 steps, a fixed covariance matrix \mathbf{C}_0 is used. After this, the algorithm learns the scale and orientation for better proposals based on past samples and generates proposals using the updated (2×2) covariance matrix in

$$\mathbf{C}_n^* = s_D \text{Cov}(\boldsymbol{\theta}_i^{(0)}, \dots, \boldsymbol{\theta}_i^{(n)}) + s_D \varepsilon \mathbf{I}_D, \quad (4.31)$$

where \mathbf{I}_D is the (2×2) identity matrix and ε is a very small nugget to prevent \mathbf{C}_n^* becoming singular. A standard scaling factor $s_D = 2.4^2/D$ is calculated based on the dimensionality

of the proposal $D = 2$. To further accelerate the AMH algorithm by avoiding large matrix inversions, a recursive calculation for Eq. 4.31 is used, as suggested by [36].

For the parameters $\log(\eta_i)$ and $\log(\nu_i)$ in the LGM, a different technique is used where MH samples are proposed using a nearby full conditional distribution that can be derived.

$$\log\left(\frac{\mathcal{K}}{Q} - 1\right) = \log(\eta) - \nu \log(t), \quad (4.32)$$

leading to the slightly different likelihood function for a well i of

$$\begin{aligned} p(\tilde{\mathbf{Q}}_i | \log(\eta_i), \nu_i, \mathcal{K}_i, \mathbf{t}_i) \\ = \prod_{m=1}^{M_i} \mathcal{N}\left(\log\left(\frac{\mathcal{K}_i}{Q_{im}} - 1\right); \log(\eta_i) - \nu_i \log(t_{im}), \sigma_{\tilde{Q}}^2\right). \end{aligned} \quad (4.33)$$

When a Gaussian prior is assumed for ν_i and $\log(\eta_i)$ this results in an alternative posterior distribution that allows for their full conditionals to be calculated. This alternative posterior distribution is similar but not equivalent to the desired target posterior distribution. As a result, it can only be used as a good proposal distribution for a metropolis-Hastings step which accepts or rejects these proposals according to Eq. 4.22, where π is still the true posterior distribution in Eq. 4.21 and the alternative posterior distribution

$$\begin{aligned} \tilde{\pi}(\log(\mathcal{K}_i), \log(\eta_i), \nu_i | \mathbf{t}_i, \mathbf{Q}_i) \\ \propto p(\tilde{\mathbf{Q}}_i | \log(\mathcal{K}_i), \log(\eta_i), \nu_i, \mathbf{t}_i) p(\log(\mathcal{K}_i)) p(\log(\eta_i)) p(\nu_i) \end{aligned} \quad (4.34)$$

is used as the proposal distribution q . This approach is similar to an independence sampler since proposals do not depend on the previous state of the parameter being sampled, however they do depend on the state of other parameters. It is also similar to a Gibbs sampler but uses a nearby full-conditional distribution so the acceptance ratio is used to ensure the sampling is adjusted to match the target. The noise parameter $\sigma_{\tilde{Q}}^2$ is chosen to be larger than σ_Q^2 to ensure good coverage of the true posterior.

The proposal distribution for $\log(\eta_i)'$ is found by deriving the full conditional for the alternative posterior distribution which uses the likelihood distribution in Eq. 4.33 and a

Gaussian prior on $\log(\eta_i)$, as in

$$\begin{aligned}
p(\log(\eta_i) | \{\mathbf{t}_i\}_{i=1}^W, \{\mathbf{Q}_i\}_{i=1}^W, \{\Psi_i\}_{i=1}^W, \{\log(\mathcal{K}_i)\}_{i=1}^W, \{\nu_i\}_{i=1}^W, \{\zeta_j\}_{j=1}^J) \\
\propto p(\mathbf{Q}_i | \log(\eta_i), \nu_i, \log(\mathcal{K}_i), \mathbf{t}_i) p(\log(\eta_i)).
\end{aligned} \tag{4.35}$$

The associated log-conditional posterior distribution is

$$\begin{aligned}
& \log [p(\mathbf{Q}_i | \log(\eta_i), \nu_i, \log(\mathcal{K}_i), \mathbf{t}_i) p(\log(\eta_i))] \\
& \propto -\frac{1}{2\sigma_{\tilde{Q}}^2} \sum_{m=1}^{M_i} \left(\log \left(\frac{\mathcal{K}_i}{Q_{im}} - 1 \right) - \log(\eta_i) + \nu_i \log(t_{im}) \right)^2 - \frac{[\log(\eta_i) - \mu_\eta]^2}{2\sigma_\eta^2} \\
& \quad \propto \left(\frac{1}{\sigma_{\tilde{Q}}^2} \sum_{m=1}^{M_i} \left[\log \left(\frac{\mathcal{K}_i}{Q_{im}} - 1 \right) + \nu_i \log(t_{im}) \right] + \frac{\mu_\eta}{\sigma_\eta^2} \right) \log(\eta_i) \\
& \quad \quad \quad - \frac{1}{2} \left[\frac{M_i}{\sigma_{\tilde{Q}}^2} + \frac{1}{\sigma_\eta^2} \right] \log(\eta_i)^2.
\end{aligned} \tag{4.36}$$

From this, the precision and mean for the Gaussian distribution used to generate proposals of $\log(\eta_i)'$ are as shown in Eq. 4.37 and Eq. 4.38.

$$\frac{1}{\sigma_{\eta_i'}^2} = \tau_{\eta_i'} = \frac{M_i}{\sigma_{\tilde{Q}}^2} + \frac{1}{\sigma_\eta^2} \tag{4.37}$$

and

$$\mu_{\eta_i'} = \frac{1/\sigma_{\tilde{Q}}^2 \sum_{m=1}^{M_i} [\log(\mathcal{K}_i/Q_{im} - 1) + \nu_i \log(t_{im})] + \mu_\eta/\sigma_\eta^2}{\tau_{\eta_i'}}. \tag{4.38}$$

Similarly the full conditional distribution for ν_i' based on the alternative posterior distribution constructed using the alternative likelihood in Eq. 4.33 and a Gaussian prior on ν_i is

$$\begin{aligned}
p(\nu_i | \{\mathbf{t}_i\}_{i=1}^W, \{\mathbf{Q}_i\}_{i=1}^W, \{\Psi_i\}_{i=1}^W, \{\log(\mathcal{K}_i)\}_{i=1}^W, \{\log(\eta_i)\}_{i=1}^W, \{\zeta_j\}_{j=1}^J) \\
\propto p(\mathbf{Q}_i | \log(\eta_i), \nu_i, \log(\mathcal{K}_i), \mathbf{t}_i) p(\nu_i).
\end{aligned} \tag{4.39}$$

Note that to make the derivation of this full conditional possible, it is necessary to work with ν_i directly, rather than the log-transformation of it. As such, $\bar{\nu}$ and S_ν^2 are used to denote the prior mean and variance to distinguish these from the hyperparameters of the

log-transformed distribution. This leads to the log-conditional distribution for ν_i as

$$\begin{aligned}
& \log [p(\mathbf{Q}_i | \log(\eta_i), \nu_i, \log(\mathcal{K}_i), \mathbf{t}_i) p(\nu_i)] \\
& \propto -\frac{1}{2\sigma_{\tilde{Q}}^2} \sum_{m=1}^{M_i} \left(\log \left(\frac{\mathcal{K}_i}{Q_{im}} - 1 \right) - \log(\eta_i) + \nu_i \log(t_{im}) \right)^2 - \frac{(\nu_i - \bar{\nu})^2}{2S_{\nu}^2} \\
& \propto \left(-\frac{1}{\sigma_{\tilde{Q}}^2} \sum_{m=1}^{M_i} \left[\log \left(\frac{\mathcal{K}_i}{Q_{im}} - 1 \right) - \log(\eta_i) \right] \log(t_{im}) + \frac{\bar{\nu}}{S_{\nu}^2} \right) \nu_i \\
& \quad - \frac{1}{2} \left(\frac{1}{\sigma_{\tilde{Q}}^2} \sum_{m=1}^{M_i} [\log(t_{im})]^2 + \frac{1}{S_{\nu}^2} \right) \nu_i^2. \quad (4.40)
\end{aligned}$$

From this, proposals for ν'_i can be drawn from the Gaussian distribution with parameters

$$\frac{1}{\sigma_{\nu'_i}^2} = \tau_{\nu'_i} = \left(\frac{1}{\sigma_{\tilde{Q}}^2} \sum_{m=1}^{M_i} [\log(t_{im})]^2 + \frac{1}{S_{\nu}^2} \right) \quad (4.41)$$

and

$$\mu_{\nu'_i} = \frac{-1/\sigma_{\tilde{Q}}^2 \sum_{m=1}^{M_i} [\log(\mathcal{K}_i/Q_{im} - 1) - \log(\eta_i)] \log(t_{im}) + \bar{\nu}/S_{\nu}^2}{\tau_{\nu'_i}}. \quad (4.42)$$

where the notation S_{ν}^2 and $\bar{\nu}$ are used to denote the empirical Bayesian variance and mean hyperparameters for ν to distinguish them from the hyperparameters for the Gaussian prior on $\log(\nu)$.

The entire MH within Gibbs algorithm is outlined in Algorithm 1. Subroutines are also outlined for the Gibbs update step in Algorithm 2, the MH step using the nearby full conditional distribution in Algorithm 3, and the adaptive MH step in Algorithm 4.

4.3 Application

In this section, the approach formulated in Sect. 4.2 is applied to two of the most prolific and longest producing unconventional basins in North America: the Barnett shale gas play in north Texas and the Bakken tight oil formation in North Dakota. This provides a test bed for both PCMs since the SCM is a forward model based on the typical conditions in the Barnett [84] and the more general LGM was previously shown to be appropriate for Bakken

Algorithm 1 Metropolis-Hastings within Gibbs for hierarchical model

- 1: Initialize $\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_W^{(0)}, \zeta_1^{(0)}, \dots, \zeta_J^{(0)}$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $\boldsymbol{\theta}_i^{(n)} \leftarrow \boldsymbol{\theta}_i^{(n-1)}, \forall i \in \{1, \dots, W\}$
- 4: $\zeta_j^{(n)} \leftarrow \zeta_j^{(n-1)}, \forall j \in \{1, \dots, J\}$
- 5: **for all** $j \in \{1, \dots, J\}$ **do**
- 6: Gibbs update ζ_j (Alg. 2)
- 7: **end for**
- 8: **for all** $i \in \{1, \dots, W\}$ **do**
- 9: **if** model is LGM **then**
- 10: Gibbs update $\log(\mathcal{K}_i)$ (analogous to Alg. 2)
- 11: Nearby full conditional MH step for $\log(\eta_i)$ MH step (Alg. 3)
- 12: Nearby full conditional MH step for $\log(\nu_i)$ (analogous to Alg. 3)
- 13: **end if**
- 14: **if** model is SCM **then**
- 15: Adaptive MH step for $\boldsymbol{\theta}_i$ (Alg. 4)
- 16: **end if**
- 17: **end for**
- 18: **end for**

Algorithm 2 Gibbs update ζ_j

- 1: Compute $\sigma_{\zeta'_j}^2(\boldsymbol{\theta}^{(n)}, \zeta_{\sim j}^{(n)})$ using Eq. 4.25
- 2: Compute $\mu_{\zeta'_j}(\boldsymbol{\theta}^{(n)}, \zeta_{\sim j}^{(n)})$ using Eq. 4.26
- 3: Propose $\zeta'_j \sim \mathcal{N}(\mu_{\zeta'_j}, \sigma_{\zeta'_j}^2)$
- 4: $\zeta_j^{(n)} \leftarrow \zeta'_j$

Algorithm 3 MH step for $\log(\eta_i)$ using nearby full conditional

- 1: Compute $\sigma_{\eta'_i}^2(\mathcal{K}_i^{(n)}, \nu_i^{(n)}, \zeta_j^{(n)})$ using Eq. 4.37
- 2: Compute $\mu_{\eta'_i}(\mathcal{K}_i^{(n)}, \nu_i^{(n)}, \zeta_j^{(n)})$ using Eq. 4.38
- 3: Propose $\log(\eta_i)' \sim \mathcal{N}(\mu_{\eta'_i}, \sigma_{\eta'_i}^2)$
- 4: $\boldsymbol{\theta}'_i \leftarrow \{\log(\mathcal{K}_i)^{(n)}, \log(\eta_i)', \log(\nu_i)^{(n)}\}$
- 5: Compute $\alpha(\boldsymbol{\theta}_i^{(n)}, \boldsymbol{\theta}'_i) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}'_i | \mathbf{t}_i, \mathbf{Q}_i) \bar{\pi}(\boldsymbol{\theta}_i^{(n)} | \mathbf{t}_i, \mathbf{Q}_i)}{\pi(\boldsymbol{\theta}_i^{(n)} | \mathbf{t}_i, \mathbf{Q}_i) \bar{\pi}(\boldsymbol{\theta}'_i | \mathbf{t}_i, \mathbf{Q}_i)} \right\}$ using Eq. 4.8 and Eq. 4.34
- 6: Generate $u \sim \mathcal{U}(0, 1)$
- 7: **if** $u \leq \alpha(\boldsymbol{\theta}_i^{(n)}, \boldsymbol{\theta}'_i)$ **then**
- 8: $\log(\eta_i)^{(n)} \leftarrow \log(\eta_i)'$
- 9: **end if**

Algorithm 4 Adaptive MH step for θ_i

```
1: if  $n \geq n_0$  then
2:   Compute  $C_n^*$  using Eq. 4.31
3: else
4:    $C_n^* = C_0$ 
5: end if
6: Propose  $\theta'_i \sim \mathcal{N}(\theta_i^{(n)}, C_n^*)$ 
7: Compute  $\alpha(\theta_i^{(n)}, \theta'_i) = \min \left\{ 1, \frac{\pi(\theta'_i | t_i, \mathbf{Q}_i)}{\pi(\theta_i^{(n)} | t_i, \mathbf{Q}_i)} \right\}$  using Eq. 4.8
8: Generate  $u \sim \mathcal{U}(0, 1)$ 
9: if  $u \leq \alpha(\theta_i^{(n)}, \theta'_i)$  then
10:   $\theta_i^{(n)} \leftarrow \theta'_i$ 
11: end if
```

wells [18].

4.3.1 Data

There are unique attributes to the data available for each of these producing regions. Widespread drilling activity commenced earlier in the Barnett than the Bakken shale so there are many Barnett wells with longer production histories. As a result, the focus here is on wells in the Barnett shale with at least ten years of production and on Bakken wells with at least five years of production. The analysis was confined to part of the Barnett shale primarily targeting gas (excluding an oilier area) and the middle Bakken formation located within North Dakota.

Both states require reports to be filed with the the surface and bottomhole location for wells and the total length of the producing lateral section. The calculated midpoint location is thus used to approximate the location of a well within the basin. For the Bakken, the depth of the well is also used since it varies considerably across the basin and is known to correlate strongly with other important subsurface conditions such as pore-overpressure and source rock maturity [96]. Production data must be reported to the regulator on a monthly basis in both states, but North Dakota additionally requires the number of producing days in each month to be reported. Wells often have downtime for maintenance or planned cycling of wells to meet surface facility constraints, so t was adjusted for these off-days with all Bakken

	Barnett	Bakken
Number of wells	1968	1083
Minimum production	10 years	5 years
PCM	SCM	LGM
θ_i	$\{\log(\mathcal{M}_i), \log(\tau_i)\}$	$\{\log(\mathcal{K}_i), \log(\eta_i), \log(\nu_i)\}$
KL modes	40	40
\mathbf{x}_i	Intercept, original gas in place, lateral length	Intercept, hydrocarbon pore volume, depth, lateral length, hydraulic fracturing water volume
Dimensionality	3979	3294

Table 4.1: Data and models for Barnett and Bakken

wells to remove this source of noise. FracFocus (web: fracfocus.org) also collects data on hydraulic fracturing designs for wells, and the total water used in hydraulic fracturing was thus available for many Bakken wells. This data for wells in each region was acquired from the data firm Rystad Energy (web: rystadenergy.com). As with the analysis of [84], wells with erratic and potentially misleading production behavior resulting from restimulation, in which the well is hydraulically fractured a second time to revive production, were excluded from the analysis. Based on this criteria, data from 1968 wells in the Barnett and 1083 wells in the Bakken could be included here.

In addition to the design properties and locations of wells, some important estimated geological properties have been pulled from geological maps in literature. [45] constructed a map of estimated original gas in place (OGIP) across the Barnett shale play using data from well logs and provided shapefiles for this analysis. Additionally, hydrocarbon pore volume estimates were obtained for all locations in the Bakken formation from [37], which had similarly been mapped using log data. These volumetric quantities should be closely linked to the \mathcal{M}_i and \mathcal{K}_i parameters governing the total amount of resource accessed by a well. This makes it ideal prior information to include, along with the technical factors previously identified, in the covariates \mathbf{x}_i for the mean trend. Table 4.1 summarizes the data and models used in the analysis.

4.3.2 Results

The MCMC simulation was run for 10^5 steps to allow the chains to adequately explore the high-dimensional joint posterior. The first 5000 samples of each chain were discarded as burn-in. To reduce sample autocorrelation and ease the burden of working with such a large number of samples, this was then thinned so that only every fiftieth sample in each chain was used. The results for β and the first few dimensions of ω in the Barnett shale are shown in Fig. 4-4. Only the second half of the chains are shown in order to make chain movement easier to distinguish.

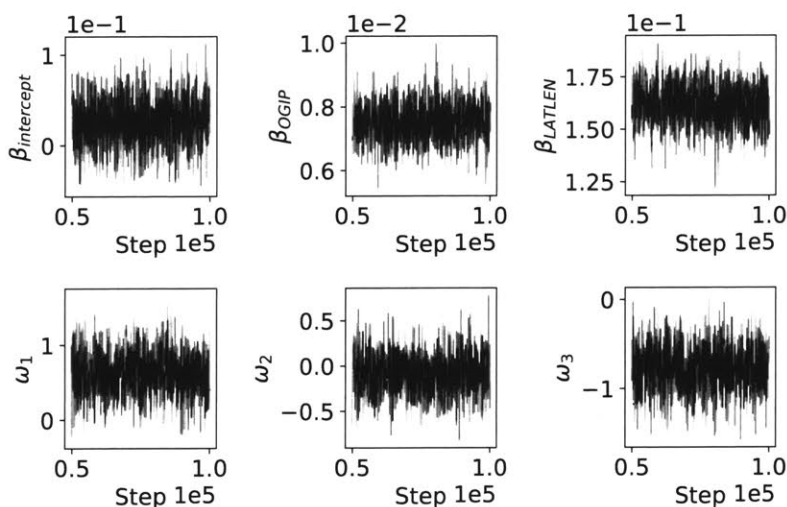


Figure 4-4: MCMC samples of coefficients for regression coefficients, including lateral length (LATLEN) and first three modes of KL expansion with Barnett wells

One benefit of a Bayesian sampling approach is that the samples generated to approximate the joint posterior distribution also provide insight into the role of individual variables in the model. Each individual parameter chain, such as those shown in Fig. 4-4, also gives the marginal distribution for that parameter. This allows the influence of physical factors such as the lateral length of wells on long-term productivity to be determined and the uncertainty of this quantified too. This can be particularly useful for modeling hypothetical scenarios and exploring different combinations of parameters. Because the hierarchical model considers the range of parameter uncertainty and other correlations within the data, estimates can differ substantially from those obtained using linear regression (LR) or regression-kriging (RK) on

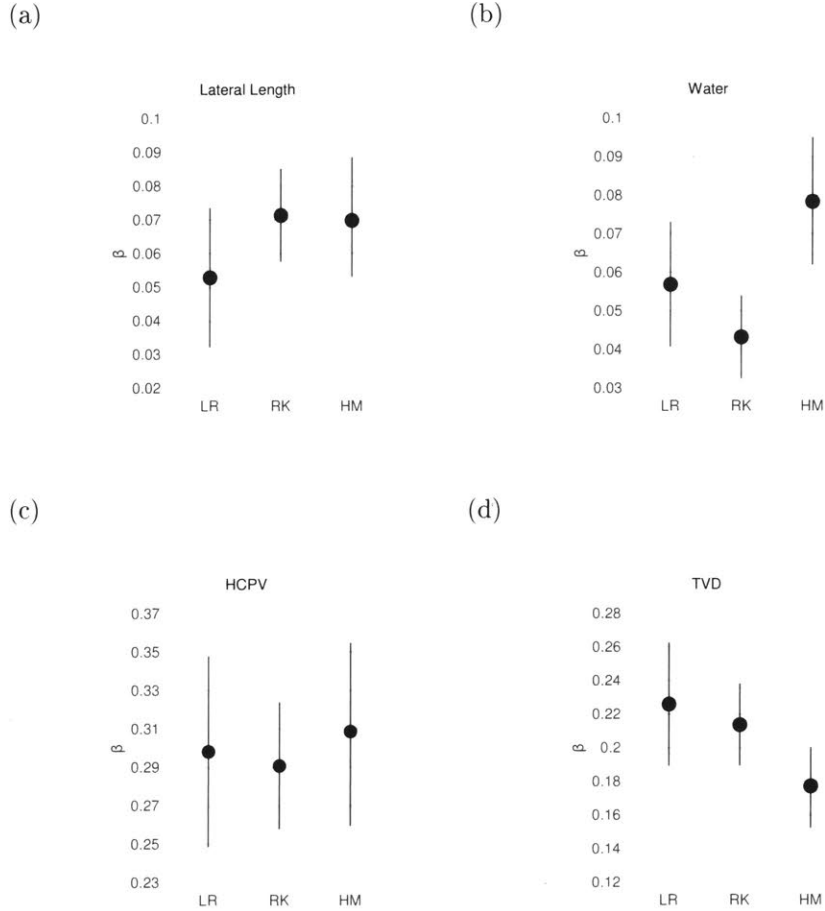


Figure 4-5: Coefficients (mean and 95% confidence or credible intervals) from linear regression (LR), regression-kriging (RK), and Hierarchical model (HM) with Bakken wells

deterministic least-squares PCM parameter estimates, as shown in Fig. 4.3.2 for properties in the Bakken shale. This reveals that hydraulic fracturing water volume potentially plays a more important role and uncertainty is greater for most parameters than suggested by the less robust methods.

Furthermore, the clear spatial hyperparameters in the model allow for geological productivity to be mapped across locations, while controlling for other factors. Figure 4-6 shows the expected long-term well productivity, or more precisely the mean of the marginalized prior for $\log(\mathcal{M})$, mapped across the Barnett shale. Because $\log(\mathcal{M})$ is linear on the components of β and ω , this can be constructed using the mean of each ζ component's marginal distribution and then multiplying by Ψ_i at each location. In order to only map the spa-

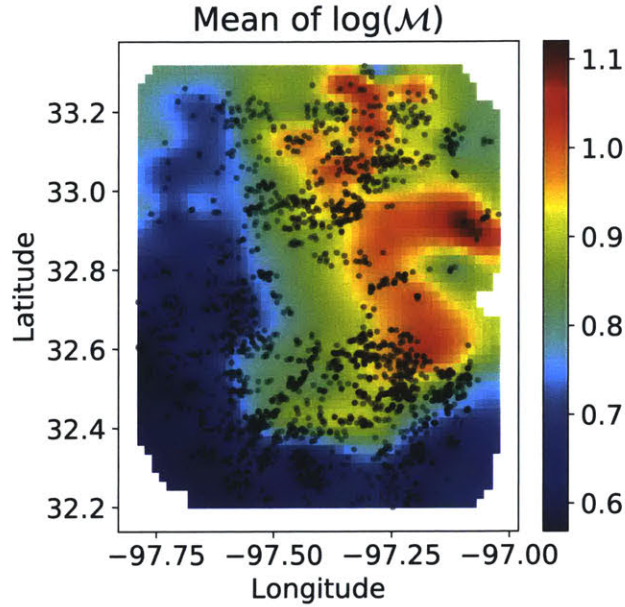


Figure 4-6: Map of wells (markers) and expected long-term well productivity in Barnett field with Lateral length held to average value

tial components, well design covariates are held constant to their mean values. Figure 4-7 similarly shows the mean of the marginalized prior for $\log(\mathcal{K})$ across the Bakken shale.

These marginal distributions provide useful insight into the system but it is important to also establish the effectiveness of the mechanistic-statistical hierarchical model at making predictions. A two-fold cross-validation scheme was used to test the accuracy of the model at forecasting long-term productivity of wells based on their first year of production. This involved randomly assigning wells to two groups that alternated roles as training and testing subsets. The prediction task applied to test wells is illustrated for a typical Barnett well in Fig. 4-8 and Bakken well in Fig. 4-9. For this well forecast, only the first year of production was observed and this was used to generate a posterior predictive distribution of future production, represented in the figures by the posterior predictive mean (PPM) and the 75% credibility interval. For test set wells, the prior incorporated the marginal distribution for each ζ_i learned from training wells. The value of the ϵ_Q noise hyperparameter was set to ensure the forecast uncertainty captured the empirical uncertainty and the actual cumulative production fell within the 75% credible interval roughly 75% of the time. Each well's cross-validated posterior predictive distribution is plotted against actual outcomes for ten-year

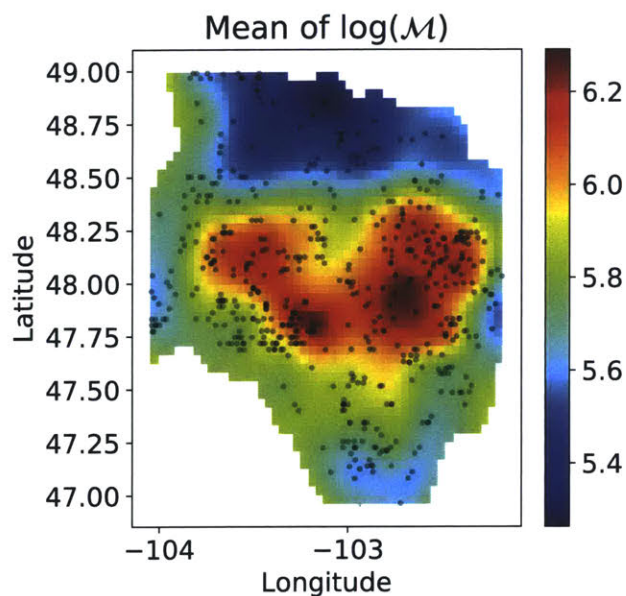


Figure 4-7: Map of wells (markers) and expected long-term well productivity in Bakken field with design parameters held to average values

cumulative production in the Barnett shale in Fig. 4-10. The same results for five-year cumulative production in the Bakken are shown in Fig. 4-11. This reveals that there is much greater uncertainty with forecasts of more productive wells, where small differences in dynamics—such as the timing of inter-fracture interference—can have dramatic implications for total production.

In order to compare the accuracy of the hierarchical method to existing deterministic forecasting techniques, the error was calculated between each well’s posterior predictive mean and actual cumulative production of each well at the chosen reference time (ten years for Barnett and five years for Bakken). The cross validation error for predictions with a standard type-well curve approach and nonlinear least squares fit of the PCM to the observed first year production was also found. Figure 4-12 shows the root mean squared error and average error over all wells in the Barnett for each method. The type-well curve and least-squares SCM in the Barnett shale give a similar RMSE but the least squares estimate systematically underestimates production. This is because without any kind of regularization, the least-squares estimate is overfitting the data and gravitating toward small parameter estimates rather than considering larger parameter values that provide a similarly good match to the limited observed production data and are also suggested by other training wells to be

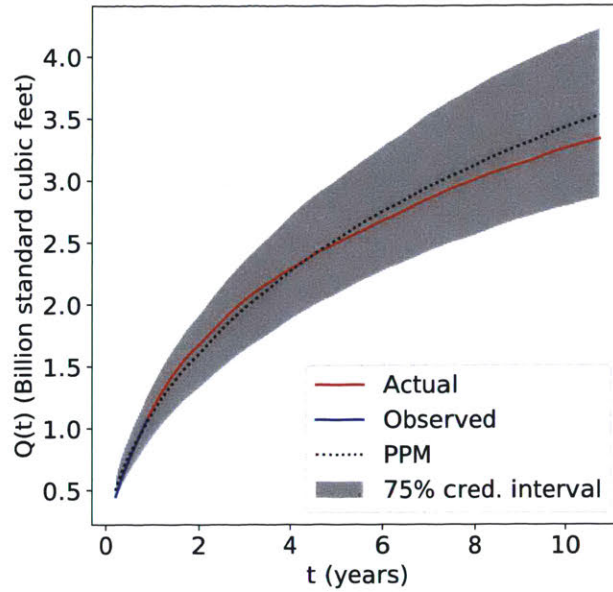


Figure 4-8: Hierarchical model fit to observed production in first year for a Barnett well giving posterior predictive mean (PPM) and credible intervals for later production

more reasonable. By contrast, the mechanistic-statistical model uses the structure of the hierarchical prior to place greater probability on parameters that better match the behavior and physical relationships observed elsewhere in the field. In the Bakken, Fig. 4-13 shows that least squares with the LGM performs slightly better than the type-well curve although it has a slight tendency to overestimate production. In both fields, the hierarchical model performs substantially better in terms of RMSE while also reducing the systematic error of least-squares predictions.

Another clear advantage of the hierarchical Bayesian approach is that it rigorously quantifies uncertainty and can potentially reduce it by including additional information in forecasts. Uncertainty with the type-well curve can be characterized by considering the full empirical distribution of offset training wells. However, this tends to overestimate uncertainty since no additional information is used to weight the relevance of these offset wells. Figure 4-14 and Fig. 4-15 show the distributions of 75% credible interval ranges for all wells in the two-fold cross-validation test compared to the implied 75% confidence interval of the type-well curve.

The random assignment of wells to training and test sets in the cross validation scheme is a very standard approach for validating a model’s predictive capability. However, it masks an important strength of the mechanistic-statistical approach that becomes clear in a more

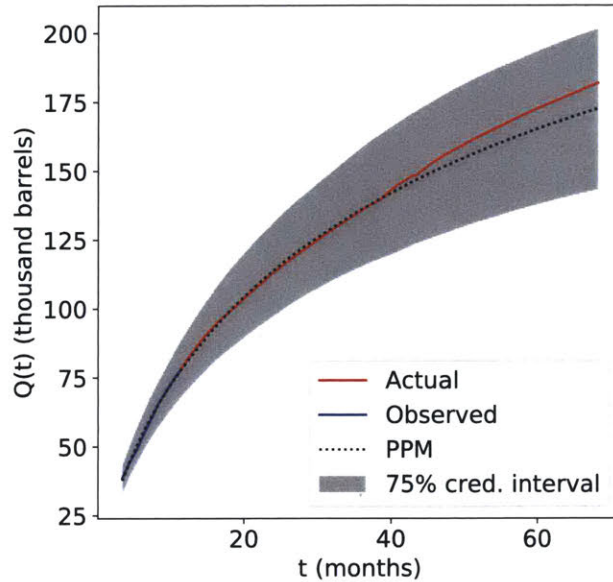


Figure 4-9: Hierarchical model fit to observed production in first year for a Bakken well giving posterior predictive mean (PPM) and credible intervals for later production

realistic forecasting scenario where wells being forecast do not match the training wells. To demonstrate this, two additional validation schemes were developed.

In the first, Bakken wells were divided into a training and test set of wells based on the volume of hydraulic fracturing water used, as shown in Fig. 4-16. This is intended to reflect the challenge today of developing forecasts for newer wells that are more intensively hydraulically fractured than their older peers in the same field, with unclear implications for production dynamics [109]. Since this test set of wells tend to be more productive than the rest of the population, the magnitude of error for all methods is greater than in random cross-validation but the relative changes are informative. Type-well curves particularly suffer in accuracy and become overly pessimistic because the training wells are no longer comparable to test wells and tend to be less productive. The hierarchical approach performs well in this situation, reducing squared error and nearly eliminating any systematic bias in forecasts by using physical relationships to adjust expectations.

Next, an even starker division was used for training and test wells in the Barnett, as shown in Fig. 4-18. The shortest 20% of wells were used for training while test wells were selected to have lateral lengths larger than 90% of wells and estimated OGIP higher than 50% of wells in the overall population. This forecasting scenario is designed to reflect the

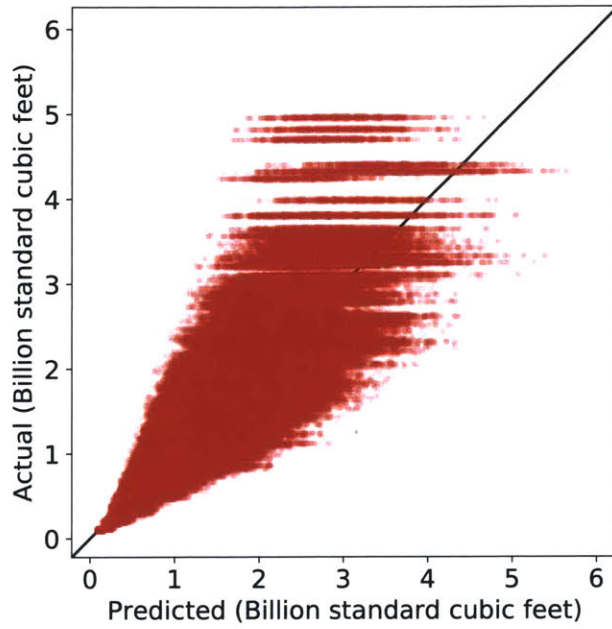


Figure 4-10: Two-fold cross validation results of hierarchical model posterior predictive and actual ten-year cumulative production for Barnett wells

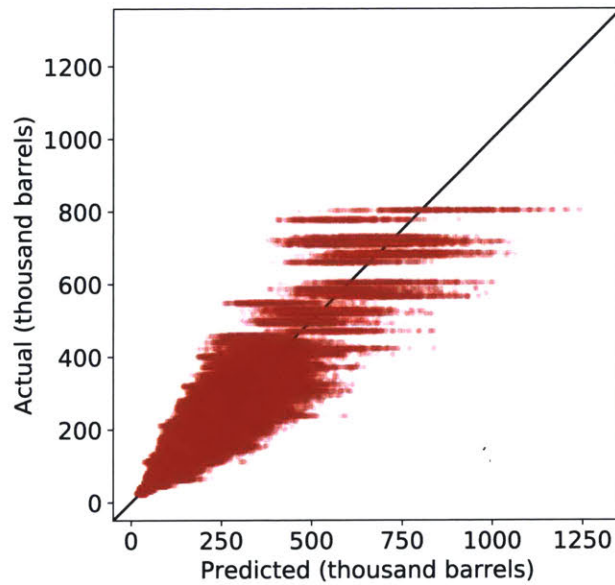


Figure 4-11: Two-fold cross validation results of hierarchical model posterior predictive and actual five-year cumulative production for Bakken wells

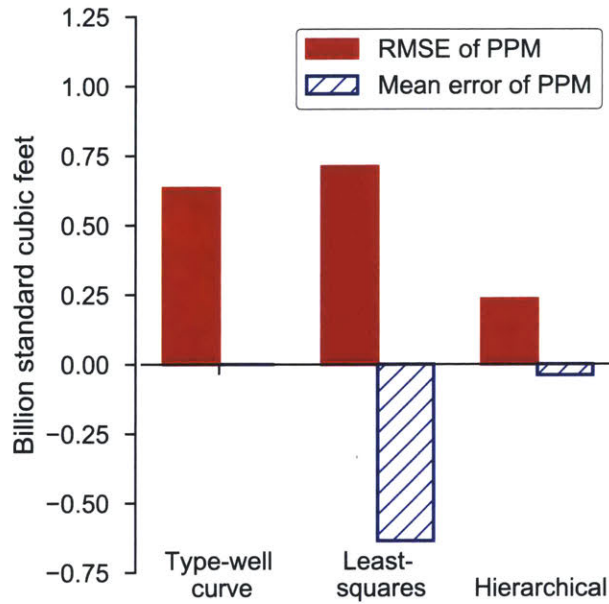


Figure 4-12: Root mean squared error (RMSE) and average error for two-fold cross validation predictions of Barnett wells' ten-year cumulative production

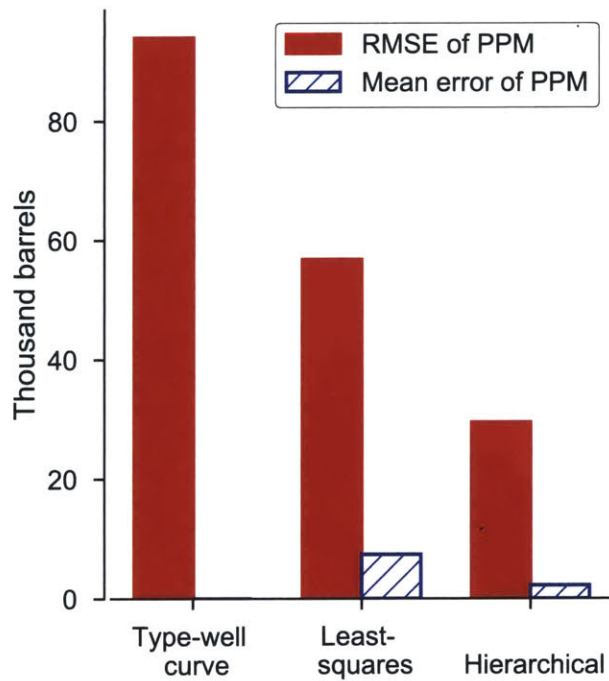


Figure 4-13: Root mean squared error (RMSE) and average error for two-fold cross validation predictions of Bakken wells' five-year cumulative production

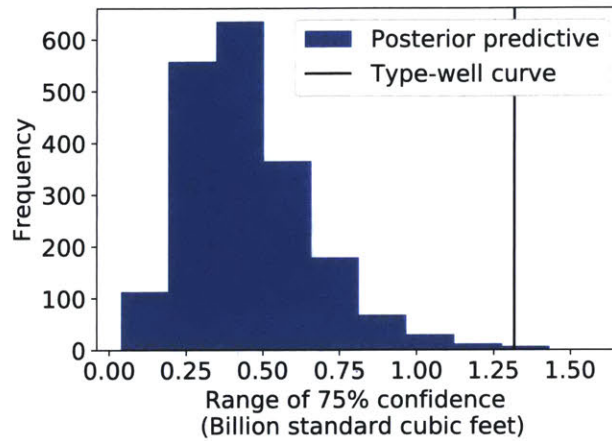


Figure 4-14: Distribution of 75% credible interval ranges for Barnett wells' ten-year cumulative production in two-fold cross validation using hierarchical model and range of 75% confidence with type-well curve

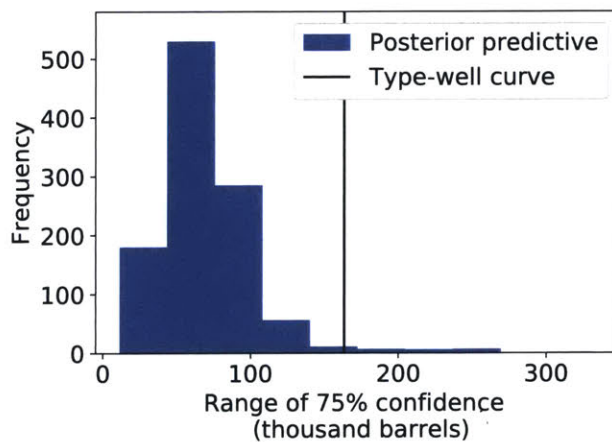


Figure 4-15: Distribution of 75% credible interval ranges for Bakken wells' five-year cumulative production in two-fold cross validation using hierarchical model and range of 75% confidence with type-well curve

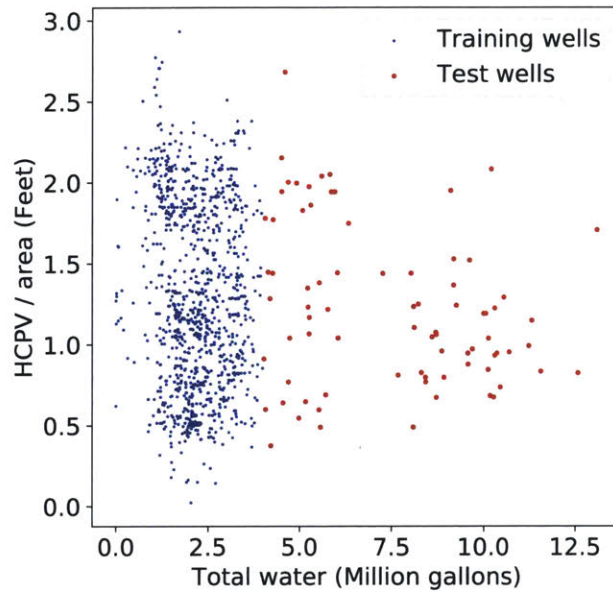


Figure 4-16: Training and test wells in Bakken shale chosen based on hydraulic fracturing water volume to test generalizability of mechanistic-statistical relationships in hierarchical model

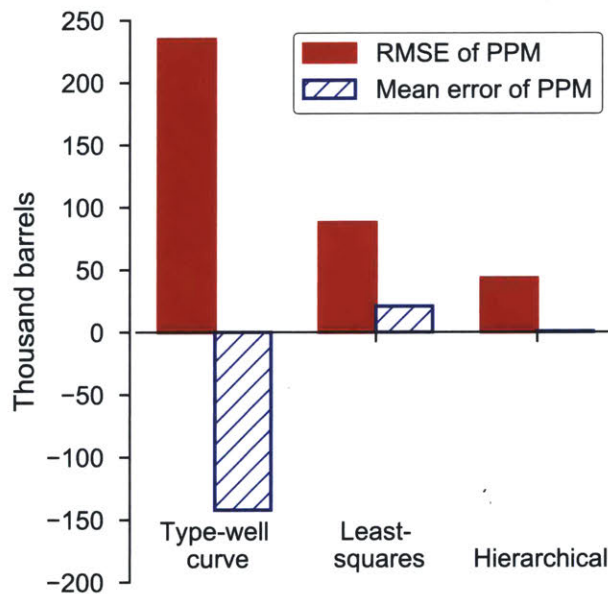


Figure 4-17: Root mean squared error (RMSE) and average error of predicted five-year cumulative production for Bakken wells with largest stimulations after learning relationship from much lower-stimulation wells

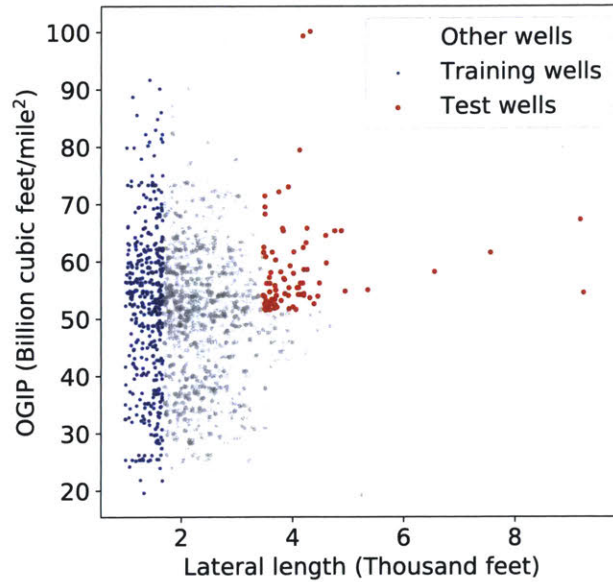


Figure 4-18: Training and test wells in Barnett shale chosen based on lateral length and reservoir quality to test generalizability of mechanistic-statistical relationships in hierarchical model

common situation early in a field’s development, when there are relatively few wells scattered throughout the field and these are much smaller appraisal wells. Once the best areas in a field are identified, activity tends to concentrate there as well designs are scaled up to maximize production from these locations [74]. The results in Fig. 4-19 show that, even with a much smaller training set of wells and a large gap in the design space, the hierarchical model is able to effectively generalize mechanistic-statistical relationships and drastically outperform other simpler methods in its accuracy. Again, the type-well curve systematically underestimates production in this realistic forecasting scenario.

The mechanistic-statistical hierarchical model is both informative about relationships in the system and accurate at forecasting. However, an empirical Bayesian approach was discussed in Sect. 4.2.2 as an approximation that eliminates some complexity and model hierarchy but still offers the benefits of uncertainty quantification and regularization. The error that this empirical Bayesian approximation introduces is examined for the different validation scenarios in Fig. 4-20. The relative change in the RMSE and mean error is shown scaled to the RMSE of the mechanistic-statistical hierarchical model. Note that a drop in the relative mean error does not necessarily indicate an improvement since it may lead

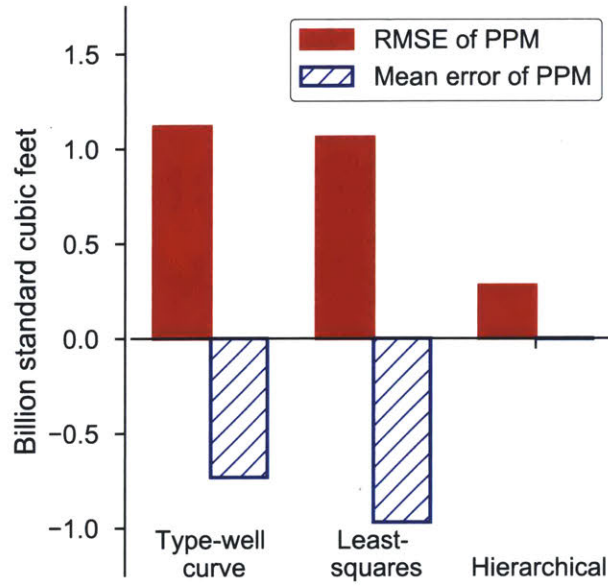


Figure 4-19: Root mean squared error (RMSE) and average error of predicted ten-year cumulative production for Barnett wells with longest laterals and in highest-quality reservoir after learning relationships from shortest wells

to greater systematic underestimation. In the unique situation where the training data is representative of the test data—as with two-fold cross validation—there is very little error introduced by using the simpler empirical Bayesian formulation. When this is a reasonable assumption to make, it may be appropriate to rely on the empirical Bayesian approximation. But when the properties of wells being forecast differ from those available as training data, as is often the case, the empirical Bayesian formulation becomes unreliable since the prior is fixed. The mechanistic-statistical approach should be used in these situations since it can adjust the prior for wells based on their physical properties.

4.4 Conclusions

This chapter demonstrated how physical data and prior knowledge about production behavior can be incorporated into a mechanistic-statistical Bayesian formulation of unconventional well production. This is a general approach, suitable in both an oil field and a gas field, and was developed here for both a physics-derived and empirical production model. By connecting wells in a field using hierarchical priors, the approach substantially improves

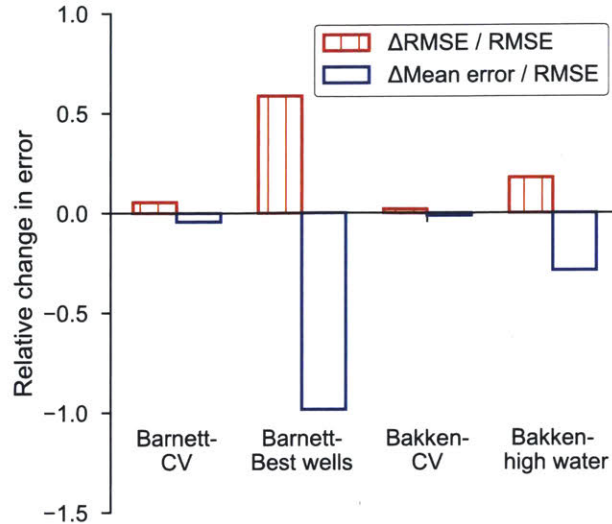


Figure 4-20: Relative change in error by using empirical Bayesian approximation instead of mechanistic-statistical hierarchical model

forecast accuracy compared to current widely used techniques of nonlinear least squares regression—often called decline curve analysis—and aggregated type-well curves. It is also better able to characterize and potentially reduce uncertainty.

The fundamental challenge in unconventional production forecasting today is appropriately applying insights from older wells with long production histories to newer wells with limited production data but known physical differences, due to changes in design and subsurface conditions. The mechanistic-statistical approach is ideally suited to this task, since it is able to learn physical relationships from older wells and use the hierarchical model structure to generalize this knowledge to wells with fundamentally different properties. This is not possible with nonhierarchical Bayesian production forecasting approaches that handle each well individually and are unable to incorporate physical data. Even the empirical Bayesian approach, discussed in this chapter as a possible approximation, lacks the probabilistic structure to detect these relationships and is unsuited for the realistic challenges of production forecasting today.

In recent years, many companies have experimented with more intensive well designs, such as larger hydraulic fracturing water volumes to increase SRV and boost production rates [103, 76]. This trend drastically increases the cost and environmental intensity of wells but the impact on long-term production behavior—and as a result, on well economics—remains

unclear. This makes it important to identify shifts in production dynamics for these wells as early as possible in order to inform design decisions for new wells and improve field-scale forecasts. Past efforts to understand the relationship of PCM parameters to known attributes of wells, including using kriging estimates in [113] and response surface modeling in [81], have neglected to consider the immense uncertainty inherent in the PCM parameters and are thus unreliable. The interpretability of the mechanistic-statistical approach makes it an important methodological step toward better understanding the drivers of long-term well productivity in unconventional fields.

The analysis here showed that current forecasting techniques have a tendency in some cases to systematically under or overestimate long-term well productivity. When applied across a large region to inform policy and infrastructure investment decisions, this can have potentially dire societal costs. Underestimation can contribute to inadequate infrastructure investment, leading to future pipeline bottlenecks and excessive flaring of stranded natural gas—an environmental bane currently common in the Permian basin of west Texas [82]. Overestimation of future production can lead to wasteful investments in economically disappointing projects and a future energy system that is potentially more expensive, environmentally damaging, and less reliable than anticipated.

It is worthwhile to contrast the use here of PCMs with a contrasting but useful idea in literature. Principal component analysis has been promoted as an approach for describing relationships of physical properties with well productivity over time [58]. This approach finds a lower dimensional subspace that maximizes variance and best represents the data. It thus provides a useful data reduction technique without assuming any model or structure. Principal component analysis (PCA) has been applied to production time series using a uniform discretization over all data [58], and in an infinite-dimensional setting by approximating this space using a finite number of basis functions [70]. This functional PCA approach was additionally combined with kriging of basis coefficients [70], which has similarities to the orthogonal Karhunen-Loève representation of the Gaussian process prior used here. Both of these previous applications of PCA to production data found essentially two dominant directions though, suggesting a simple parametric PCM is adequate to represent these dynamics. By projecting data onto an a priori manifold based on a PCM, rather than a linear

hyperplane like in PCA, patterns in the data can be more easily recognized and interpreted. This is especially true when the parameter space is defined by physical attributes, as is the case with the mechanistic-statistical approach here. PCA is unable to incorporate physics and domain knowledge and has a tendency to amplify noise without the assistance of these assumed dynamical relationships. PCA is purely descriptive, whereas the mechanistic relationships in the hierarchical model allow for generalization or extrapolation onto physically different data not previously observed, making it very practical for real world forecasting situations. Additionally, these past implementations of PCA also provided no uncertainty quantification.

A hierarchical approach is by definition flexible and additional hierarchy should be added to incorporate richer datasets and provide insight on more specific aspects of production behavior. Production was modeled here as either single-stream gas or oil production, based on the primary targeted resource in each field. However, in reality unconventional well production is multi-phase and includes both oil and gas, as well as sizable amounts of water. Ratios between these fluids evolve over time and are important to production dynamics. A natural extension of the hierarchical model would be to use three separate PCMs for each well to represent these fluids, with parameters dependent on shared, well-level priors. This would provide insight into how these fluid ratios change over time in wells [110]. Additionally, when data for well flowing pressure is available, more detailed physics models can replace the PCMs used here. Rate transient analysis is an area that attempts to do this deterministically today [19], but would be much more appropriately handled in a probabilistic hierarchical framework given the unresolvable physical uncertainties. Finally, well-to-well production interference is of growing concern as unconventional fields are drilled more densely and wells begin to compete with each other for the same resource [45]. The ability of the mechanistic-statistical approach to infer SRV properties of wells across a field makes it an indispensable tool in understanding and predicting this detrimental subsurface condition. As these potential extensions of the approach illustrate, the mechanistic-statistical model in this chapter provides a powerful and flexible framework that opens up new possibilities for using data to understand and manage these complex unconventional resources.

Chapter 5

Conclusion

5.1 Summary

This thesis developed a novel mechanistic-statistical approach to unconventional oil and gas production forecasting that improves on current approaches by more effectively integrating well-level data with physics and domain knowledge. This leads to greater accuracy in predictions, insight into the mechanisms driving production, and better model generalizability with the ability to appropriately adapt behavior observed in older wells to newer ones with physical differences and a paucity of data. The ultimate approach for this was a spatiotemporal hierarchical Bayesian model leveraging a fractured shale gas production model in Chap. 4. In order to develop this approach, some key aspects of modeling well productivity statistically using coarse public datasets had to first be addressed.

Developing reliable resource forecasts as well as economically optimizing well designs and development plans depends on the ability to isolate the influence of design improvements from heterogeneous and often unknown geology. Chapter 2 introduced regression-kriging as an improved methodology for identifying the amount of productivity improvement associated with technological changes in wells while controlling for the role of geology. This technique combines the econometric interpretability of multiple regression for well design parameters with kriging to infer unknown geological productivity of locations as a latent Gaussian process. By recognizing and learning from the structured information implicit in spatial autocorrelation of well productivity, prediction error for over 3000 wells in the Willis-

ton Basin was reduced by a third and less biased estimates of technology’s impact were obtained. This revealed that, during the period studied, roughly half of productivity improvement was driven by changes in where drilling was occurring rather than how wells were designed, a fact completely missed by the less spatially resolved approach previously relied on by the U.S. Energy Information Administration [101, 86]. Reducing this bias is critical to avoid overestimating future well production rates and potentially encouraging wasteful overly-intensive well designs.

This initial approach focused on predicting the first year of production from a well prior to being drilled—an important metric for well economics. However, there is also significant variability in temporal production behavior and it is challenging to predict based on early production data from a well. Differences in long-term dynamics become important as they are aggregated over many wells in a basin, making this a critical forecasting task for policymakers and energy sector decision-makers.

Chapter 3 drew attention to the ill-posed nature of the inverse problem typically used to forecast long-term production—fitting production curve models (PCMs) to observed production data and extrapolating the trend. It is generally recognized that this approach becomes unreliable when limited production is available, but this chapter provided the first rigorous account of the inherent model sloppiness behind this and how to address it using Tikhonov regularization. This strikes a balance between information from the individual well and patterns seen in the broader population (a form of partial-pooling), providing a forecast with half the error of the current practice of fully-pooling production data from physically dissimilar offset wells into an average type-well curve. This approach was implemented with over 4000 wells in the Barnett shale using both the widely used Arps curve and a scaling curve recently introduced as a physics-based alternative. These models are at two extremes in the debate over which PCM is most appropriate for shale wells, but this argument is moot as regularization is shown here to be far more important than choice of model. Regularization brings the accuracy of these models close to that of a deep neural network, a model-neutral benchmark introduced here as a novel way of establishing a lower bound on error by identifying the best nonlinear mapping possible with the data. Both PCMs combined with regularization approach upon this benchmark using far fewer param-

eters, reinforcing the point that what PCMs require is not greater parametric complexity but making better use of available data. Finally, the mechanistic parameters in the scaling curve provide useful insight into the physical source of forecasting uncertainty: ambiguity between the amount of resource accessed by a well and how rapidly it is being depleted.

There is a natural connection between Tikhonov regularization and the prior in a Bayesian framing of the production forecasting problem. Chapter 4 develops this Bayesian approach, which is both highly interpretable and addresses the important issue of quantifying uncertainty in forecasts and relationships. Additionally, hierarchical priors in the model allow the concepts in Chap. 2 and Chap. 3 to be effectively combined into one spatiotemporal approach. The parameters in each well's mechanistic production model are conditioned on physical data (through multiple regression) and on a Gaussian process approximated by a Karhunen-Loève expansion—essentially a more formalized version of the regression-kriging approach from Chap. 2. This partial-pooling honors the production data for each well while inferring shared relationships across all wells. This allows the regularization that was shown to be essential in Chap. 3 to be carried out in a way that automatically tunes itself to the physical differences in wells. The model is very high-dimensional, with between three and four thousand parameters for implementations here on Barnett shale gas wells and Bakken tight oil wells. A key practical innovation introduced in this chapter to address this computational hurdle was a unique Metropolis within Gibbs sampling approach for highly structured Markov chain Monte Carlo approximation. By embedding structure and domain knowledge into the problem, the mechanistic-statistical approach reduces uncertainty in forecasts and gives mean predictions with 50-80% less error than previously possible. The physical basis also makes this model better able to generalize patterns learned from older well designs to wells with more intensive designs. This is the first basin-wide spatiotemporal production model combining physics with statistics and represents a significant advance in the area of forecasting unconventional oil and gas resources.

5.2 Future work

This model opens a path for future work addressing some of the most vexing modeling challenges afflicting unconventional resource development today. The connection in this model between production rates and stimulated reservoir volume can help provide statistical insight into how well spacing, geology, and completion design contribute to well-to-well pressure communication, or interference, and “frac-hits” which can sabotage an existing parent wells production or lead to unexpectedly low production from the new child well[45, 6]. This is one of the key technical challenges of forecasting production and optimizing field development in unconventional oil and gas today; addressing it is beyond the scope of this thesis but a framework has been created which can introduce the geometric constraints from wells sited close to each other into production forecasts. Additionally, it is increasingly important to consider the influence of water, gas, and oil production from a well together. Again, the hierarchical framework introduced here can very naturally be extended to address this issue. Separate PCMs can be used to describe each fluid and their behavior linked together through a common hyperprior. This will provide a more dynamic model for multi-stream production since these ratios can evolve over time with subsurface changes. As development of unconventional oil and gas continues, there will likely be many other challenges that arise which can be addressed by building on this mechanistic-statistical framework.

Bibliography

- [1] After OPEC: American shale firms are now the oil market's swing producers. *The Economist*, 2015 May 1.
- [2] Rigonomics. *The Economist*, 2016 Jun 1.
- [3] Peering inside the Permian. *The Economist*, 2018.
- [4] Luc Anselin. Spatial econometrics. *A companion to theoretical econometrics*, pages 310–330, 2001.
- [5] J J Arps. Analysis of decline curves. *Transactions of the American Institute of Mining Engineers*, 160:228–247, 1945.
- [6] A. Awada, M. Santo, D. Lougheed, D. Xu, and C. Virues. Is that interference? A work flow for identifying and analyzing communication through hydraulic fractures in a multiwell pad. *SPE Journal*, 21(5):1554–1566, 2016.
- [7] Jason David Baihly, Raphael Mark Altman, and Isaac Aviles. Has the Economic Stage Count Been Reached in the Bakken Shale? In *SPE Hydrocarbon Economics and Evaluation Symposium*, Calgary, Alberta, 2012.
- [8] Richard A Berk. *Regression analysis*. Sage Publications, Inc., 2004.
- [9] L Mark Berliner. Physical-statistical modeling in geophysics. *Journal of Geophysical Research*, 108, 2003.
- [10] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, New York, 2011.
- [11] Roger Bivand. *Creating Neighbours*. 2016.
- [12] J Bocquet-Appel and R Sokal. *Spatial Autocorrelation Analysis of Trend Residuals in Biological Data*, 1989.
- [13] A Brandt, G Heath, E Kort, F O'Sullivan, G Petron, S M Jordaan, P Tans, J Wilcox, A M Gopstein, D Arent, S Wofsy, N J Brown, R Bradley, G D Stucky, D Eardley, and R Harriss. Methane leaks from North American natural gas systems. *Science*, 343:733–735, 2014.

- [14] Adam R. Brandt, Tim Yeskoo, and Kouros Vafi. Net energy analysis of Bakken crude oil production using a well-level engineering-based model. *Energy*, 93:2191–2198, 2015.
- [15] John Browning, Svetlana Ikonnikova, Gürcan Gülen, and Scott Tinker. Barnett Shale Production Outlook. *SPE Economics & Management*, 5(03):89–104, jul 2013.
- [16] Jenny Brynjarsdottir and L. Mark Berliner. Bayesian Hierarchical Modeling for Temperature Reconstruction From Geothermal Data. *The Annals of Applied Statistics*, 5(2B):1328–1359, 2011.
- [17] Sergio Centurion, Randall Cade, and Xin Luo. Eagle Ford Shale: Hydraulic Fracturing, Completion, and Production Trends: Part II. In *SPE Annual Technical Conference and Exhibition*, San Antonio, TX, 2012.
- [18] Aaron James Clark, Larry Wayne Lake, and Tadeusz Wiktor Patzek. Production Forecasting with Logistic Growth Models. In *SPE ATCE*, number 144790, Denver, CO, 2011.
- [19] C. R. Clarkson. Production data analysis of unconventional gas wells: Review of theory and best practices. *International Journal of Coal Geology*, 109-110:101–146, 2013.
- [20] C R Clarkson, J L Jensen, and S Chipperfield. Unconventional gas reservoir evaluation: What do we have to consider? *Journal of Natural Gas Science and Engineering*, 8:9–33, 2012.
- [21] Troy Cook and Dana Van Wagener. Improving Well Productivity Based Modeling with the Incorporation of Geologic Dependencies. 2014.
- [22] Thomas R. Covert. Experiential and Social Learning in Firms: The Case of Hydraulic Fracturing in the Bakken Shale. 2014.
- [23] Rafael Wanderley De Holanda, Eduardo Gildin, and Peter P. Valkó. Combining physics, statistics, and heuristics in the decline-curve analysis of large data sets in unconventional reservoirs. *SPE Reservoir Evaluation and Engineering*, 21(3):683–702, 2018.
- [24] Anna Driver and Terry Wade. U.S. oil output slide looms as shale firms hit productivity wall. *Reuters*, 2015 Oct., oct.
- [25] Jean Dubé and Diègo Legros. *Spatial Econometrics Using Microdata*. ISTE & Wiley, London and Hoboken, NJ, 2014.
- [26] Carsten F. Dormann, Jana M. McPherson, Miguel B. Araújo, Roger Bivand, Janine Bolliger, Gudrun Carl, Richard G. Davies, Alexandre Hirzel, Walter Jetz, W. Daniel Kissling, Ingolf Kühn, Ralf Ohlemüller, Pedro R. Peres-Neto, Björn Reineking, Boris Schröder, Frank M. Schurr, and Robert Wilson. Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, 30(5):609–628, 2007.

- [27] Timothy Fitzgerald. Experiential Gains with a New Technology: An Empirical Investigation of Hydraulic Fracturing. *Agricultural and Resource Economics Review*, 44(2):83–105, 2015.
- [28] David S Fulford, Braden Bowie, Michael E Berry, Bo Bowen, Apache Corporation, Derrick W Turk, and Terminus Data Science. Machine Learning as a Reliable Technology for Evaluating Time / Rate Performance of Unconventional Wells. (January), 2016.
- [29] Attila Gábor and Julio R. Banga. Robust and efficient parameter estimation in dynamic models of biological systems. *BMC Systems Biology*, 9(1), 2015.
- [30] Andrew Gelman, John B Carlin, Hal S Stern, David B. Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 3 edition, 2014.
- [31] Arthur Getis and Jared Aldstadt. Constructing the Spatial Weights Matrix Using a Local Statistic. *Geographical Analysis*, 36(2):90–104, 2004.
- [32] Xinglai Gong, Raul Gonzalez, Duane A. McVay, and Jeffery D. Hart. Bayesian probabilistic decline-curve analysis reliably quantifies uncertainty in shale-well-production forecasts. *SPE Journal*, 19(06):1047–1057, 2014.
- [33] Pierre Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997.
- [34] Rickey Green and Kendall Manning. Nine Energy stimulates 50-stage Divert-A-Frac system in Williston basin. *World Oil*, 237(6), 2016.
- [35] G Gullickson, Kirk Fiscus, and Peter Cook. Completion Influence on Production Decline in the Bakken/Three Forks Play. In *SPE Western North American and Rocky Mountain Joint Regional Meeting*, Denver, CO, 2014.
- [36] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.
- [37] H Scott Hamlin, Katie Smye, Robin Dommissie, Ray Eastwood, Casee R Lemons, Guinevere Mcdaid, and Economic Geology. Geology and Petrophysics of the Bakken Unconventional Petroleum System. *Unconventional Resources Technology Conference*, pages 1–14, 2017.
- [38] W.K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.
- [39] Tomislav Hengl, Gerard B M Heuvelink, and Alfred Stein. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1-2):75–93, 2004.

- [40] Keith Holdaway. *Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data-Driven Models.* chapter 8-9. John Wiley & Sons, Inc., Hoboken, 2014.
- [41] Aojie Hong, Reidar B. Bratvold, Larry W. Lake, and Leopoldo M. Ruiz Maraggi. Integrating model uncertainty in probabilistic decline-curve analysis for unconventional-oil-production forecasting. *SPE Reservoir Evaluation & Engineering*, (August 2018):23–25, 2019.
- [42] M. K. Hubbert. Nuclear energy and the fossil fuel. In *Drilling and production practice.* American Petroleum Institute, 1956.
- [43] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.
- [44] S Ikonnikova, J Browning, G Gülen, K Smye, and SW Tinker. Factors influencing shale gas production forecasting: Empirical studies of Barnett, Fayetteville, Haynesville, and Marcellus Shale plays. *Economics of Energy & Environmental Policy*, 4(1):19–35, jan 2015.
- [45] Svetlana Ikonnikova, John Browning, Susan Christine Horvath, and Scott Tinker. Well Recovery, Drainage Area, and Future Drill-Well Inventory: Empirical Study of the Barnett Shale Gas Play. *SPE Reservoir Evaluation & Engineering*, 17(04):484–496, nov 2014.
- [46] Svetlana Ikonnikova and Gürcan Gülen. Impact of low prices on shale gas production strategies. *The Energy Journal*, 36(01):43–62, 2015.
- [47] Svetlana Ikonnikova, Gürcan Gülen, John Browning, and Scott W. Tinker. Profitability of shale gas drilling: A case study of the Fayetteville shale play. *Energy*, 81:382–393, 2015.
- [48] Svetlana Ikonnikova, Katie Smye, John Browning, and Robin Domisse. Update and enhancement of shale gas outlooks. Technical report, Bureau of Economic Geology, 2018.
- [49] D. Ilk, T. A. Blasingame, and O. Houzé. Practical considerations for well performance analysis and forecasting in shale plays. *Open Petroleum Engineering Journal*, 9:107–136, 2016.
- [50] Mason Inman. The fracking fallacy. *Nature*, 516(7529):28–30, 2014.
- [51] Ghazal Izadi, Jean Junca, and Randall Cade. Multidisciplinary Study of Hydraulic Fracturing in the Marcellus Shale. In *48th US Rock Mechanics/Geomechanics Symposium*, Minneapolis, MN, 2014. American Rock Mechanics Association.
- [52] Hadi Jabbari and Zhengwen Zeng. Hydraulic Fracturing Design for Horizontal Wells in the Bakken Formation. In *46th US Rock Mechanics / Geomechanics Symposium*, Chicago, IL, 2012. American Rock Mechanics Association.

- [53] R Jackson, A Vengosh, J Carey, R Davies, T Darrah, F O’Sullivan, and G Petron. The environmental costs and benefits of fracking. *Annual review of Environment and Resources*, 2014.
- [54] D. M. Jarvie. Components and processes affecting producibility and commerciality of shale resource systems. *Geologica Acta*, 12(4):307–325, 2014.
- [55] Tor A Johansen. On Tikhonov regularization, bias and variance in nonlinear system identification. *Automatica*, 33(3), 1997.
- [56] Mark J. Kaiser. Profitability assessment of Haynesville shale gas wells. *Energy*, 38(1):315–330, 2012.
- [57] Mark J Kaiser. Haynesville Louisiana Drilling and Production Update 2012. *Natural Resources Research*, 24(1):1–19, 2014.
- [58] Aaditya Khanal, Mohammad Khoshghadam, W. John Lee, and Michael Nikolaou. New forecasting method for liquid rich shale gas condensate reservoirs with data driven approach using principal component analysis. *Journal of Natural Gas Science and Engineering*, 2017.
- [59] Yuji Kim and Nori Nakata. Geophysical inversion versus machine learning in inverse problems. *The Leading Edge*, 37(12):894–901, 2018.
- [60] Randy F LaFollette, William D Holcomb, and Jorge Aragon. Impact of Completion System, Staging, and Hydraulic Fracturing Trends in the Bakken Formation of the Eastern Williston Basin. In *SPE Hydraulic Fracturing Technology Conference*, 2012.
- [61] J Lee and R Sidle. Gas-reserve estimation in resource plays. *Society of Petroleum Engineers Economics & Management*, 2:86–91, 2010.
- [62] Julie A LeFever. Isopach of the Three Forks Formation (GI-64). Technical report, North Dakota Geological Survey, 2008.
- [63] Julie A LeFever. Structural Contour and Isopach Maps of the Bakken Formation in North Dakota (GI-59). Technical report, North Dakota Geological Survey, 2008.
- [64] JP LeSage and RK Pace. *Introduction to spatial econometrics*. Chapman and Hall/CRC, 2009.
- [65] Noam Lior. Exergy, Energy, and Gas Flow Analysis of Hydrofractured Shale Gas Extraction. *Journal of Energy Resources Technology*, 138(6), 2016.
- [66] E Lolon, K Hamidieh, L Weijers, M Mayerhofer, H Melcher, and O Oduba. Evaluating the Relationship Between Well Parameters and Production Using Multivariate Statistical Models : A Middle Bakken and Three Forks Introduction to Statistical Modeling. In *SPE Hydraulic Fracturing Technology Conference*, The Woodlands, Texas, 2016.

- [67] Youssef M. Marzouk and Habib N. Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862–1902, 2009.
- [68] Christophe McGlade, Jamie Speirs, and Steve Sorrell. Methods of estimating shale gas resources – Comparison, evaluation and implications. *Energy*, 59:116–125, 2013.
- [69] Christophe McGlade, Jamie Speirs, and Steve Sorrell. Methods of estimating shale gas resources: Comparison, evaluation and implications. *Energy*, 59:116–125, 2013.
- [70] Alessandra Menafoglio, Ognjen Grujic, and Jef Caers. Universal Kriging of functional data: Trace-variography vs cross-variography? Application to gas forecasting in unconventional shales. *Spatial Statistics*, 15:39–55, 2016.
- [71] Gregory Meyer and Nathalie Thomas. National Grid and NY governor tussle over energy supplies, sep 2019.
- [72] Russell B. Millar and Renate Meyer. Non-linear state space modelling of fisheries biomass dynamics by using Metropolis-Hastings within-Gibbs sampling. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 49(3):327–342, 2000.
- [73] Shahab Mohaghegh, Khalid Mohamad, Andrei Popa, Sam Ameri, and Dan Wood. Performance Drivers in Restimulation of Gas-Storage Wells. *SPE Reservoir Evaluation & Engineering*, 4(06):536–542, 2001.
- [74] J.B. Montgomery and F.M. O’Sullivan. Spatial variability of tight oil well productivity and the impact of technology. *Applied Energy*, 195:344–355, 2017.
- [75] Justin Montgomery. Invited lecture at Energy Forecasting Forum. In *US Department of Energy*, Washington, DC, 2018.
- [76] Justin Montgomery and Francis O’Sullivan. A Statistical Framework for Data-Driven Assessment of Unconventional Oil and Gas Resources. In Martin J. Clifford, Robert K. Perrons, Saleem H. Ali, and Tim A. Grice, editors, *Extracting Innovations: Mining, Energy, and Technological Change in the Digital Age*, chapter 11, pages 145–161. Taylor & Francis, Boca Raton, FL, 2018.
- [77] Justin B Montgomery. *Characterizing Shale Gas and Tight Oil Drilling and Production Performance Variability*. Massachusetts Institute of Technology, Engineering Systems Division, Cambridge, Massachusetts, 2015.
- [78] Dan Murtaugh. Shale Drillers Pump More Oil From Each Well as Rigs Mean Less. *Bloomberg News*, 2015 Sept.
- [79] Shirley Neff and Margaret Coleman. EIA outlook: Reversal in U.S. oil import dependency. *Energy Strategy Reviews*, 5:6–13, 2014.
- [80] I.O.A. Odeh, A.B. McBratney, and D.J. Chittleborough. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma*, 67(3-4):215–226, 1995.

- [81] B. A. Ogunyomi, S. Dong, N. La, L. W. Lake, and C. S. Kabir. An approach to modeling production decline in unconventional reservoirs. *Journal of Petroleum Exploration and Production Technology*, 8(3):871–886, 2018.
- [82] Oil & Gas Journal. Permian gas flaring, venting reaches record high, jun 2019.
- [83] B Olson, R Elliott, and CM Matthews. Fracking’s secret problem-oil wells aren’t producing as much as forecast. *The Wall Street Journal*, nov 2018.
- [84] Tad W Patzek, Frank Male, and Michael Marder. Gas production in the Barnett Shale obeys a simple scaling theory. *Proceedings of the National Academy of Science*, 110(49):19731–19736, 2013.
- [85] Richard M. Pollastro, Laura N. R. Roberts, and Troy A. Cook. Geologic Model for the Assessment of Technically Recoverable Oil in the Devonian-Mississippian Bakken Formation, Williston Basin. *Shale reservoirs-Giant resources for the 21st century*, 97:205–257, 2012.
- [86] Jim Polson and Tim Loh. U.S. Vastly Overstates Oil Output Forecasts, MIT Study Suggests. *Bloomberg Businessweek*, 2017.
- [87] W. Press, S.A. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 2 edition, 1992.
- [88] Rystad Energy. Rystad Energy Shale Commentary-Halliburton outperforms the competition in Bakken: mixed empirical evidence. Technical report, 2017.
- [89] Luigi Saputelli, Carlos Lopez, Alejandro Chacon, and Mohamed Soliman. Design Optimization of Horizontal Wells With Multiple Hydraulic Fractures in the Bakken Shale. In *SPE/EAGE European Unconventional Resources Conference and Exhibition*, Vienna, Austria, 2014.
- [90] Jared Schuetter, Srikanta Mishra, Battelle Memorial, Ming Zhong, Baker Hughes, Randy Lafollette, and Baker Hughes. Data Analytics for Production Optimization in Unconventional Reservoirs. 2015.
- [91] Robert Shelley, Nijat Guliyev, and Amir Nejad. A Novel Method to Optimize Horizontal Bakken Completions in a Factory Mode Development Program. In *SPE Annual Technical Conference and Exhibition*, San Antonio, TX, 2012.
- [92] Haipeng Shen and Zhengyuan Zhu. Efficient mean estimation in log-normal linear models. *Journal of Statistical Planning and Inference*, 138(3):552–567, 2008.
- [93] SPEE. Guidelines for application of the Petroleum Resources Management System. Technical report, 2011.
- [94] A M Stuart. Inverse problems : A Bayesian perspective Inverse problems. *Acta Numerica*, 19:451–559, 2010.

- [95] Lei Tan, Lihua Zuo, and Binbin Wang. Methods of decline curve analysis for shale gas reservoirs. *Energies*, 11(3), 2018.
- [96] Cosima Theloy. Integration of Geological and Technological Factors Influencing Production in the Bakken Play, Williston Basin. pages 151–156, 2014.
- [97] Wei Tian, Jitian Song, and Zhanyong Li. Spatial regression analysis of domestic energy in urban areas. *Energy*, 76:629–640, 2014.
- [98] Mark K Transtrum, Benjamin B Machta, and James P Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*, 83.3(036701):1–35, 2011.
- [99] A Tsoularis and J Wallace. Analysis of logistic growth models. *Mathematical Biosciences*, 179:21–55, 2002.
- [100] U.S. Energy Information Administration. Recent drilling activity has lowered costs and increased performance. Technical report, U.S. Energy Information Administration, Washington, DC.
- [101] U.S. Energy Information Administration. Annual Energy Outlook 2014. Technical report, Washington, DC, 2014.
- [102] U.S. Energy Information Administration. Drilling Productivity Report. Technical report, 2016.
- [103] U.S. Energy Information Administration. Trends in U. S. Oil and Natural Gas Upstream Costs. Technical report, Washington, DC, 2016.
- [104] U.S. Energy Information Administration. Annual energy outlook 2018 with projections to 2050. Technical report, US EIA, Washington, DC, 2018.
- [105] U.S. Energy Information Administration. Oil and gas supply module of the National Energy Modeling System. Technical report, US EIA, Washington, DC, 2018.
- [106] Fabián Vera, Casee Lemons, Ming Zhong, William D Holcomb, Randy F Lafollette, and Baker Hughes. Multidisciplinary Approach in the Permian Basin - A Geological , Statistical and Engineering Case Study to Production Results on the Wichita-Albany. In *SPE Hydraulic Fracturing Technology Conference*, The Woodlands, Texas, 2015.
- [107] P. F. Verhulst. Mathematical Researches Into the Law of Population Growth Increase. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18:1–42, 1845.
- [108] M.C. Vincent. The Next Opportunity To Improve Hydraulic-Fracture Stimulation. *Journal of Petroleum Technology*, 64(03):118–127, 2012.
- [109] HanYi Wang. What Factors Control Shale-Gas Production and Production-Decline Trend in Fractured Systems: A Comprehensive Analysis and Investigation. *SPE Journal*, 22(02):562–581, 2017.

- [110] Lei Wang, Shihao Wang, Ronglei Zhang, Cong Wang, Yi Xiong, Xishen Zheng, Shangru Li, Kai Jin, and Zhenhua Rui. Review of multi-scale and multi-physical simulation technologies for shale and tight gas reservoirs. *Journal of Natural Gas Science and Engineering*, 37:560–578, 2017.
- [111] Christopher K Wikle. Hierarchical Models in Environmental Science. *International Statistical Review*, 71(2):181–199, aug 2003.
- [112] B J a Willigers, S Begg, and R B Bratvold. Hot spot hunting: Optimising the staged development of shale plays. *Journal of Petroleum Science and Engineering*, 149:553–563, 2017.
- [113] Zhenke Xi and Eugene Morgan. Combining decline curve analysis and geostatistics to forecast gas production in the Marcellus shale. *SPE Reservoir Evaluation & Engineering*, 2019.
- [114] Wei Yu, Xiaosi Tan, Lihua Zuo, Jenn Tai Liang, Hwa C. Liang, and Suojin Wang. A new probabilistic approach for uncertainty quantification in well performance of shale gas reservoirs. *SPE Journal*, 21(6):2038–2048, 2016.
- [115] Qiumei Zhou, Robert Dilmore, Andrew Kleit, and John Yilin Wang. Evaluating gas production performances in marcellus using data mining technologies. *Journal of Natural Gas Science and Engineering*, 20:109–120, 2014.