An Exploration of and Case Studies in Demand Forecast Accuracy: Replenishment, Point of Sale, and Bounding Conditions

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Kevin Barry Smyth, M.S.

Graduate Program in Business Administration

The Ohio State University

2017

Dissertation Committee:

A. Michael Knemeyer, Advisor

Keely L. Croxton

Rod Franklin

Xiang (Sean) Wan

Abstract

Forecasts are a critical input that drive actions within the firm and throughout the supply chain. For good reason, there is a tremendous focus on accuracy for this input. This dissertation addresses three areas regarding forecast accuracy in logistics and the supply chain relating to three questions posed by demand planners at a logistics provider firm that partnered with this research. In attempting to determine "What is causing our replenishment forecast error?", "What predictive factors can help improve our demand forecast accuracy?", and with regards to forecast accuracy "How good is good enough?", we explore three interrelated topics that have a broader impact on the academic conception of forecast accuracy than the original questions posed.

In three essays, we identify governance form factors that affect replenishment forecast deviation and bias, demonstrate accuracy improvement though the inclusion of uncertain weather forecast information in demand forecasts, and identify themes that serve to bound achievable and desirable demand forecast accuracy through a systematic literature review of logistics and supply chain journals. Our first study measures the deviation and bias related to franchise governance form, but also demonstrates a novel approach to contextualize the heterogeneity of effects across regionally, temporally and product category related conditions. Our second study expands on previous work linking the inclusion of uncertain weather forecast variables to improvements in demand forecast

accuracy by examining a wider range of products and locations in a new industry, but also by demonstrating the limits to the value of uncertain information. Finally, our systematic literature review comprehensively presents the current state of research on the thematic drivers of forecast accuracy.

Each essay expands theoretical understanding of management phenomena, and reframes the manner in which previous research can be applied in practice. In each we also propose future avenues to expand on the work here, and on forecasting in general in the context of logistic and the supply chain.

Acknowledgments

I first wish to thank my committee and, and especially my advisor for guiding me through this process. Their advice, constructive criticism, and at times moral support helped get me through a long and arduous effort. Drs. Mike Knemeyer, Keely Croxton, Rod Franklin and Xiang Wan all dedicated a tremendous amount of time and energy to ensure the success of my research. By their support, this product is superior to what would be possible through my work alone.

I want to thank the supply chain planning division of our practitioner company research partner for their patience and support as we worked with them on multiple projects. To Scott Saunders and Chris Karsten for facilitating our research through the Global Supply Chain Forum at The Ohio State University. To Chris Awalt for his continued effort to guide, inform and provide for the collaboration between our research team and their firm. To Jagadeesh Deva, Brent Labhart, Jim O'Rourke and Caroline Francisca for their guidance and facilitation of data collection. Without all of their efforts, none of this research would have been possible, and it is my sincere hope that this work contributes to their success as well.

Thanks to Dr. Douglas Lambert and the all those that help run the Global Supply Chain Forum. Their work facilitating a collaborative research environment for

practitioners and academics served as the genesis of this research, and enabled the required resources to see it through.

Finally, and most importantly, I want to thank my wife Emily. Her love and support gave me the strength to complete this project. She sacrificed her time and shouldered additional burdens with our family, all while excelling as the communications director at the Virtual Labs School at The Ohio State University and pursuing a Master's of Social Work at Indiana University.

Vita

2008	B.S. Nuclear Engineering with a Minor in Aerospace Studies, Oregon State University
2012	M.S. Logistics and Supply Chain Management, Air Force Institute of Technology (AFIT)
2012	Graduate Certificate, Nuclear Weapon Effects, Proliferation and Policy, AFIT
2008 to present	Munitions and Missile Maintenance Officer, United States Air Force
2014 to present	Graduate Teaching Associate, Department of Marketing and Logistics, The Ohio State University

Fields of Study

Major Field: Business Administration

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The origin of this research is a collaboration with a fourth party logistics (4PL) provider for a large multinational quick service restaurant firm. Through the Global Supply Chain Forum at The Ohio State University, we met with the 4PL and discussed research opportunities. The firm had been experiencing three related issues in their demand planning. After discussing their concerns, we determined that their issues coincided with under-investigated themes in logistics research all relating to demand forecast accuracy. These three issues were then transformed into three sets of research questions with the aim of creating a better understanding of demand forecast accuracy. The following research is divided into three essays. Each essay relates to a distinct, but related aspect of forecast accuracy.

The first essay stems from the first question posed by the 4PL: "What is causing our replenishment forecast error?" This question relates to the way that the restaurant firm generated replenishment forecasts. The 4PL first collaborates with their restaurant firm client to generate a demand forecast for each menu item at each restaurant. This occurs each week on a rolling basis. Each restaurant location is given the estimate for demand for the upcoming week, and is responsible to purchase the raw materials required to meet the predicted demand.

To inform this process, the 4PL generates a forecast for bi-weekly replenishment derived from their own demand forecasts. This estimate includes safety stock and waste factors peculiar to each restaurant location. For instance, lead time from a distributor and available storage space may affect how much can be stored at one time, and how much may be discarded at each restaurant. By policy, each restaurant receives a supply shipment twice each week from one of the firm's distribution partners. The individual restaurant, as the entity most affected by errors in a replenishment forecast and presumably in the best position to understand local conditions not captured in a statistical forecast, is given the final say on replenishment orders.

The problem that arises is that while restaurant order edits may help the individual restaurant, they introduce variance upstream in the supply chain. Distributors, who also receive predicted replenishment orders from the 4PL, have to react to changes in replenishment orders initiated at the restaurant level in order to meet contractually mandated service levels. This increase in variance increases required levels of safety stock at distribution centers, as most often distributors have already placed orders with their suppliers by the time restaurant replenishment orders are made. This is costly not only because inventory and warehousing costs increase, but as the products are largely perishable, increased stock levels also increase product loss to spoilage.

In our first essay, we examine the potential reasons why restaurants edit their orders in a data exploration case study. We frame the investigation around the theme of agency costs arising from governance form. Previous works (Brickley and Dark 1987, Norton 1988a, Bertagnoli 1989, Krueger 1990, Carney and Gedajlovic 1991, Kaufmann

and Lafontaine 1994, Michael 2000, Yin and Zajac 2004, de Leeuw, Holweg and Williams 2011) indicate franchise owners are more likely than corporate outlet managers to act independently of the best interests of the parent franchisor firm, and consistent with local incentives. Profit motivated franchise owners are also more likely than revenue motivated corporate outlet managers not to shirk, more actively monitor costs, and waste less (Rubin 1978, Krueger 1990, Noren 1990, Norton 1988a, Norton 1988b).

This first work, titled "The Effect of Governance Form on Replenishment Forecast Deviation and Bias: A Case Study in the Quick-Service Restaurant Industry", promises to answer the 4PL firm's question of what may be driving their replenishment forecast error. In it, we simultaneously address a deficiency in management literature on the post-contractual operational effects of utilizing the franchise governance form.

The second question posed by the 4PL, and the genesis of our second essay, is: "What predictive factors can help improve our demand forecast accuracy?" The firm was eager to incorporate additional information into their demand forecasts, but were unsure of what information was valid to include, and what effect it may have on their forecast accuracy.

The firm's demand forecasts, as described above, are generated once each week and cover two replenishment periods. To be predictive, information must be available to the decision maker prior to a decision being made. This means that demand planners at the 4PL must have knowledge of events up to and including one week in the future in order to effectively include the information in a demand forecast.

We frame our second essay around the inclusion of uncertain information in a demand forecast, and our prediction for if and how consumers will respond to these external factors. For predictive information to hold value, it must be sufficiently reliable over a forecast horizon, and managers must be able to make material changes as a result of the information (Thompson and Brier 1955, Thompson 1962, Murphy 1977, Katz and Murphy 1990, Katz and Lazo 2011). One readily available source of predicted future information is that of short term weather. Weather sensitivity research, though well established in fields like agriculture and mining, is lacking in restaurant and service industries (Lazo et al. 2011, Bujisic et al. 2016). Work that incorporates weather forecasts into demand forecasts is similarly sparse, but also currently extant only in industries unrelated to the focal firm.

In our second essay, titled: "The Impact of Including Forward Indicators on POS Demand Forecast Accuracy: The Case of Short-Term Weather Forecast Data", we measure the effect of including predicted weather on demand forecast accuracy. We predict demand effects are driven by consumer mood and utility (Steele 1951, Starr-McCluer 2000, Tran 2016), and that this effect is heterogeneous across regions and product categories.

This not only answers the 4PL's question of whether this particular set of forward indicators can improve their demand forecasts, but also addresses two deficiencies in management literature. Namely, it more systematically answers what effect does weather have on demand in the restaurant industry, and what are the positive or deleterious effects of using uncertain weather information to generate demand forecasts.

The third question from the 4PL firm, “How good is good enough?”, is the origin of our third essay. By “good enough”, they mean that they wish to know when the pursuit of greater accuracy in demand and replenishment forecasts is a lost cause. We recognized that there does not appear to be any work that addresses this question holistically, and endeavored to map out a current understanding of all potential bounds for demand forecast accuracy.

To this end, we sought to answer this question through a systematic review of logistics and supply chain management literature. In our third essay, titled: “A Literature Review and Typology of Factors that Bound Demand Forecast Accuracy”, we explore relevant supply chain and logistics academic journals in an attempt to identify a typology of factors that form tradeoffs in forecast accuracy.

Relevant search topics aim to identify conditions that drive requirements and capabilities for greater forecast accuracy, but also situations when it is only possible or indeed desirable to accept lower levels of forecast accuracy. As the voluminous literature on demand forecasting can be quite diverse in focus, it was imperative that we properly scoped our search to those topics most relevant to a logistics or supply chain setting. General conditions for consideration included the effects of error signal amplification through a supply chain, cost tradeoffs of information gathering and processing, level of aggregation, impact of manual forecast adjustments, effect of information sharing, effect of accuracy metrics used, effects of overfitting, degree of supply chain integration and collaboration, length of lead time and forecast horizon, the impact of product substitution, supply chain flexibility and resilience, and effect of demand variability. Together, this

body of literature on tradeoffs in various dimensions of forecast accuracy is formed into a general typology. Such a framework can be used to inform future research defining individual tradeoffs as well as the interplay between multiple factors, and the identified types that require further investigation for full understanding are identified. This typology also serves as a self-assessment tool for the 4PL, addressing their third question.

In answering the three questions posed by our 4PL research partner: "What is causing our replenishment forecast error?", "What predictive factors can help improve our demand forecast accuracy?", and "How good is good enough?", we also fill shortfalls that currently exist in management literature in three dimensions. Our first essay explores factors that drive error in upstream forecasts, framed around governance form effects on post-contractual performance. Our second essay examines the effect of including uncertain predictive factors in demand forecasts, framed on information uncertainty and heterogeneous consumer response. Our final essay reviews the current academic understanding of factors that form trade-offs in demand forecast accuracy, and proposes a typology for future research development. All three essays include forecast accuracy as their focal construct, and contribute to a better understanding of factors that affect it.

Chapter 2: "The Effect of Governance Form on Replenishment Forecast Deviation and Bias: A Case Study in the Quick-Service Restaurant Industry"

Introduction

The strategic decision to outsource the rights to use one's business model and brand to an outside party through franchise agreements is one that is made by many different types of firms. As of the 2012 census, franchisors account for 12.1% (560,086) of all establishments and 17.4% ($1.454 Trillion) of all revenue among U.S.-based businesses (Census Bureau 2012). As a result, researchers have developed a body of literature focused on the antecedents of and causes for the managerial decision to pursue this business approach. In the process they have identified firm, market, and situational characteristics that significantly contribute to the ex-ante development of franchise governance form. However to date, an examination of the ex-post supply chain performance characteristics of firms pursuing this business model is lacking.

Franchise agreements are developed with the intent of maximizing realized long-term profits to a parent firm. Management literature to date has identified multiple sources of potential drivers for pursuing a franchise business model, and the vast majority of this research has been dedicated to identifying the significant antecedents of a firm's decision to implement this organizational form. But how might the use of the franchise governance form influence the supply chain performance of the firm after a contract is in place? In their comprehensive review of recent management literature on franchising,

Combs et al. (2010) note a dearth of "operations management" franchising research. Specifically, they identify no articles exploring how parent firms can create value for their franchisee clients (and vice versa) via supply chain activities. de Leeuw et al. (2011) examined inventory performance in franchised business units in the automotive industry. While they did not directly compare forms of governance, they did observe differential performance between distributed outlets with decentralized inventory control. The heterogeneity stemmed from incentives unique to each outlet. As a result, they call for research comparing inventory policies in devolved supply chains, which most notably includes firms employing the franchise governance form.

In an effort to understand the operational supply chain implications of pursuing a franchise model, we measure whether governance form has a significant effect on replenishment forecast deviation and bias beyond what can be explained by other factors. We examine rates of deviation and bias in the replenishment forecast of a focal firm that utilizes both company-owned and franchised outlets, and discuss the ramifications of governance form with respect to outlet inventory and replenishment policy. To accomplish these research goals, we collected distribution center level replenishment forecast and inventory data from a major U.S.-based restaurant chain that utilizes company-owned and franchised retail outlets. The analysis will identify whether franchised outlets experience higher levels of replenishment forecast deviation and whether this deviation has a greater negative bias than their company-owned counterparts, when controlling for other factors. We will observe whether these results are consistent with extant management literature on governance form as a result of

agency costs. Theoretical and managerial implications of these findings will then be described.

Research Context

The origins of this study stem from an ongoing relationship with corporate members of a university-affiliated research group. Through the group, we started working with the primary fourth party logistics (4PL) provider for a major international quick service restaurant to develop research questions that simultaneously address relevant theoretical management issues and solve multiple supply chain issues their major restaurant customer was experiencing. One such problem was significant order deviation by individual restaurant outlets in their centrally developed replenishment forecasts. Managers and owners at individual restaurant locations are permitted the freedom to deviate from centrally planned order levels due to the likelihood that they would have greater knowledge of local conditions that would drive demand. While this policy is likely to improve operations at the outlet level, deviations drive uncertainty in replenishment and are viewed as error by distributors.

The restaurant firm in our research operates almost 15,000 retail outlets domestically, of which more than 80% are contracted to franchise companies. The firm utilizes the aforementioned 4PL to centrally develop sales forecasts and plan replenishment for all restaurant outlets, and five third party logistics (3PL) companies provide warehousing and distribution services for the restaurant chain. The restaurant chain manages national and regional promotions as part of their demand planning process, but each outlet may add local promotional activity that tends to be underreported

and can impact centralized forecast accuracy. All involved firms, i.e. the restaurant firm, its 4PL provider, its 3PL providers, and the franchise companies, utilize a management information system (MIS) curated by the 4PL. Each entity's relevant MIS data are visible to at least the adjacent link in the supply chain. Figure 1 illustrates the relevant data flows within and between firms.

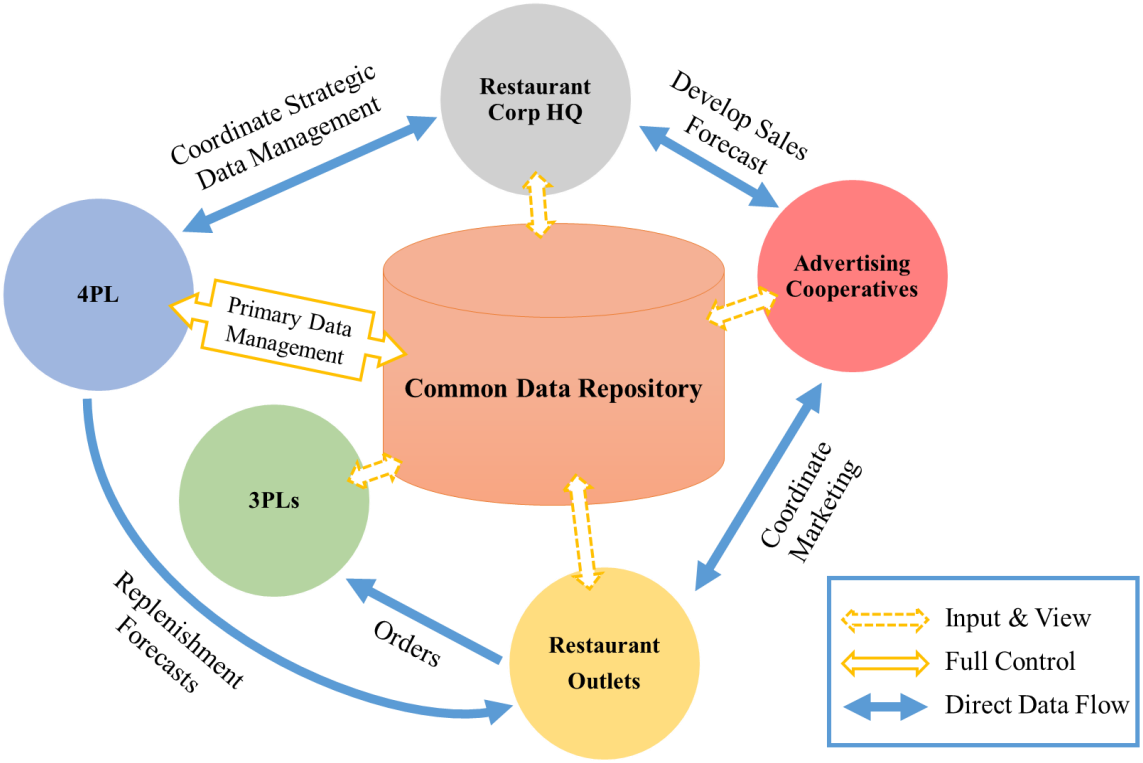

Figure 1: Restaurant Firm Data Flows

With this complex replenishment process involving so many parties, we had to determine our best course of action to identify potential causes of deviation in the replenishment forecast. Interviews within the 4PL firm, and with one 3PL provider

indicated that demand information asymmetry (Lee et al. 2000, Cachon and Fisher 2000) was not likely a cause of deviation, as multiple echelons share data that are updated daily. Nor were delays in information (Chen 1999), as MIS data are updated daily reflecting both distribution center (DC) and restaurant inventory levels, as well as the projected effect of daily orders. Anecdotally, data entry errors, particularly at the individual restaurant level introduce some noise in the replenishment process and may lead to deviation, but this did not appear to be a primary source. Interviews with the restaurant firm's promotions team, as well as with one 3PL distribution team, indicated restaurant governance form could be a source of deviation. Specifically, these teams perceived franchised outlets to exhibit higher levels of deviation and a more positive bias in their deviations than the restaurant company-owned outlets. The restaurant firm's financial disclosures corroborated this account by indicating outlet governance form was a significant operational risk factor, as franchised locations were seen as more likely to deviate from corporate standard procedure.

The logic in permitting both franchised and company-owned outlets the discretion to edit the centrally developed replenishment forecasts is that outlet-level management are more likely to have an understanding of local demand conditions (including locally planned promotions) beyond what is captured in the statistical forecast. This notion is heavily supported in forecasting literature (Fildes et al. 2008, Syntetos, Boylan and Disney 2009, Eroglu and Knemeyer 2010, Eroglu and Croxton, 2010), and subjective adjustments are often incorporated in the demand planning process. However, any order from the outlet that differs from the forecast replenishment amount constitutes a deviation

that must be accounted for with buffer stock at the DC (Zinn et al. 1989, Gardner 1990). Each 3PL-operated DC receives a recommended buffer stock level from the 4PL planning firm, but has the discretion to establish their own stock levels. These DC operators are contractually obligated to meet service level requirements of all company-owned and franchised outlets, and are responsible for any charges incurred from stock-outs and emergency shipments. The result of deviation by the restaurant outlet operator is that the restaurant firm's supply chain holds excessive system-wide inventory across echelons. In addition to incurring unnecessary holding costs, many stock keeping units are perishable or tied to a finite promotion, so some portion of excess inventory will constitute an irrecoverable loss. DC operators will inevitably pass these costs on to their supply chain partners.

Replenishment forecasts are currently developed in the same way regardless of governance form. While this may lower the cost and complexity of demand management for the firm, management literature on governance form suggests that franchisees may behave differently when it comes to replenishment, which contributes to additional inventory carrying costs at the next higher echelon in the supply chain (de Leeuw et al. 2011). As residual claimants of their local business unit, franchisees are more likely than their firm-employed managerial counterpart to behave in a manner that maximizes their local interests (Lafontaine 1992, Michael 2000, Kaufmann and Lafontaine 1994, Shelton 1967, Krueger 1991, Brickley and Dark 1987, Rubin 1978). To date, studies have used aggregated factors to describe ex-ante determinants of organizational form. This research

is positioned to identify specific ex-post performance effects of the choice of organizational form.

Literature Review

In order to understand the franchising governance form and the performance implications of its adoption, we first examine extant literature on governance form. Many possible theoretical explanations for the strategic decision to pursue a franchise approach have been posited; most prominently transaction costs, resource scarcity and agency costs. We review each briefly, but derive our primary theoretical motivation from agency theory.

Transaction Costs

Transaction cost theorists see franchisor firms as an interstice or hybrid of pure vertical integration and the open market (Rubin 1978, Norton 1988a, Norton 1988b). Caves and Murphy II (1976) describe franchising as an "intracorporate management of decentralized units". Ex-ante investment in a franchise by an entity legally separate from the firm acts as a "hostage" in Williamsonian terms, bridging the gap between a pure market and internal corporate transactions (Noren 1990). While most authors agree that this hybrid form exists, the argument for causal mechanisms that drive the level of firm-like or market-like characteristics of this organization form, as well as the extent to which a firm chooses to franchise their outlets tends to split between a resource based view and an administrative efficiency view.

Resource Scarcity

The idea that a firm chooses to franchise its outlets as a result of a lack of access to resources, which are then supplied by franchisees, was originated by Oxenfeldt and Kelly (1969), and further asserted by Caves and Murphy II (1976). Rooted in the resource based theory of the firm, franchising allows for firm growth even when resources for expansion are scarce. Rapid growth is viewed as desirable, as it allows a firm to build brand capital, take advantage of a market opportunity to seize a larger share of total demand, or achieve economies of scale in promotion and production (Carney and Gedajlovic 1991). Growth of a firm permits further acquisition and dynamic replenishment of the available pool of productive resources (Penrose 1959), and is limited primarily by the market, capital resources and managerial resources.

Due to the problem of adverse selection, in which lenders have inadequate cues to evaluate the availability of alternatives, Combs and Ketchen (1999) assert franchising can alleviate the problem of resource scarcity by giving firms access to capital that can be less costly than capital from debt and equity markets. Firms may also access additional entrepreneurial capacity by changing investors into managers. Franchising transfers increased residual ownership and risk to managers, eliminating the need for a loan or capital market investment while controlling net monitoring (or those associated with observing and controlling actions of outlet managers) costs (Norton 1988a). Carney and Gedajlovic (1991) and Mathewson and Winter (1985) both support this resource scarcity view, but in an effect moderated by administrative efficiency and stage of firm development.

However, multiple studies refute the logical development of a resource scarcity based motivation for franchising. Franchising as a capital scarcity expedient is countered with the proposition that passive investment is diversified over all of a firm's outlets, presumably demanding a lower return on this lower risk (Rubin 1978, Brickley and Dark 1987, Lafontaine 1992). Detractors found either no significant effect from scarcity on organizational form, or found that effect to be dominated or heavily moderated by the effect from agency costs (Rubin 1978, Brickley and Dark 1987). In a meta-analysis of 44 previous studies on the causes of governance form choice, Combs and Ketchen (2003) conclusively redirect the argument, finding no significant resource scarcity drivers for the choice of organizational form. They did, however, find numerous significant agency costs that may help explain the choice to franchise.

Agency Costs or Administrative Efficiency

For this research, the most relevant theoretical explanation for the determination of governance form is agency theory, which describes the two-sided moral hazard that exists between principals (firm owners) and agents (firm employees). In the context of this research, this hazard exists between the parent restaurant firm central planners and outlet owners or managers. Both work toward their own self-interest where possible, but agency loss can be mitigated through bonding by the agent and monitoring by the principal. Franchising is a hybrid of these two control mechanisms (Brickley and Dark 1987).

Franchise bonding occurs by transferring residual risk and claim to a legally (though not practically) separate firm (Rubin 1978). This is typically in the form of an

ex-ante investment by the franchisee, usually as a franchising fee and a non-returnable real investment, who in turn receives excess rents as the principal of their own firm. Bonding may also include elements beyond simple claims on expected profit; such as supernormal ex-ante and ex-post rents left unclaimed in franchise contracts (Kaufmann and Lafontaine 1994), expectation for contract renewal, or promise of additional lucrative contract award (Bertagnoli 1989).

This method of ex-ante bonding investment is intended to capture all of the expected excess profits for the parent firm after discounting a reasonable return to the franchisee, monitoring costs and residual agency costs such as those related to shirking. Due to bounded rationality (or the inability of the parent firm to predict all future profits with certainty), incomplete franchise agreements (Mathewson and Winter 1985) necessitate ex-post royalties that are some percent of revenue (Caves and Murphy II 1976, Rubin 1978, Noren 1990, Combs and Ketchen 2003) in addition to the initial fee.

Bonding is a less direct means of minimizing the two-sided moral hazard, in which it is difficult or expensive for the principal and agent to observe the other's actions, by both eliminating the agent's incentive to shirk (Bertagnoli 1989, Noren 1990, Krueger 1991, Carney and Gedajlovic 1991, Michael 2000) and ensuring the principal's continued assistance and interest in the outlet's performance (Rubin 1978, Combs and Ketchen 2003).

Post contractual monitoring is the primary mechanism designed to eliminate the remaining moral hazard after bonding costs are established in the contract (Rubin 1978, Fama and Jensen 1983, Yin and Zajac 2004, Paik and Choi 2007). Monitoring refers to

continued costs related to observing and controlling agent actions after contract formation and bonding, which could include formation and dissemination of corporate policies to outlets, or travel and technology costs to observe outlet behavior. This is a more direct approach that is effective when there is direct coercive control and outlets are easily accessible. Where monitoring is relatively more expensive, such as when outlets are at greater distances from a corporate headquarters or in a market with unknown or transient conditions, bonding is substituted (Fama and Jensen 1983, Norton 1988a, Norton 1988b, Lafontaine 1992, Combs and Ketchen 2003). Monitoring costs are found to be more significant among franchised outlets (Brickley and Dark 1987, Norton 1988a, Norton 1988b, Lafontaine 1992, Combs and Ketchen 2003), but some elements of cost are mitigated in recent years by advances in telecommunications and MIS (Yin and Zajac 2004, Cochet et al. 2008).

After implementing both control mechanisms, there remains a residual loss as a result of self-interested actions that could not be controlled (Mahoney 2000, Jensen and Meckling 1976). Firms choose the franchise governance form more often for firms with high monitoring costs, and in doing so substitute bonding costs to a greater extent. Therefore, residual agency loss among franchised units is more likely to come as a result of inefficient bonding than inefficient monitoring. We should expect the manner of royalty collection between the two governance forms to reflect this difference in significance.

Managers of firm-owned outlets are typically compensated on a fixed or revenue-based scale, so additional costs from inefficiencies, shirking and perquisite taking do not

significantly affect their personal income. This revenue basis of evaluation is due to the fact that it is difficult for parent firms to attribute drivers of revenue or cost to actions of a manager independent of the firm. Franchisees, on the other hand, typically pay royalties based on revenue and claim any realized profit (Rubin 1978, Krueger 1991). This increased claim comes with increased risk. Franchisees, who may have large proportions of their personal wealth tied up in an outlet, and may be restricted in their property rights to quickly sell assets (Noren 1990), directly feel any costs from inefficiencies. Given this greater risk experienced by franchisees, they have a greater motivation than managers at corporate outlets to actively monitor their own operations. An example of this is in labor costs. Franchise governance form is positively related to cost structures with large labor components (Norton 1988a, Norton 1988b). Presumably as a result of more engaged management, franchise employees are found to report higher levels adequate managerial supervision, and outlets experience lower rates of perquisite taking, lower mean wage, and lower rate of wage increase (Krueger 1991).

Though residual loss components are more prevalent in choosing the franchise model, inefficient risk bearing, free ridership, and quasi rent appropriation drive the use of the corporate owned outlet. Increasing risk (and thereby interest and effort) of the franchisee may cause inefficient risk bearing, or the unwillingness to make optimal investments as a result of heavy undiversified investment in a single outlet. This cost is found to be more significant in choosing the franchise governance form (Brickley and Dark 1987, Carney and Gedajlovic 1991). Underinvestment may also relate to what is known as the externality or free rider problem, where the franchisee bears less risk from

underinvesting in advertising, customer experience, and overall quality (Rubin 1978, Mathewson and Winter 1985, Brickley and Dark 1987, Carney and Gedajlovic 1991, Michael 2000). This effect is found to manifest among franchisees of a parent firm with high brand value who receive a significant portion of their patronage from customers unlikely to return, as is the case with business catering to travelers (Norton 1988b), though this also has contrapositive examples (Brickley and Dark 1987). Free riding loss can be based on brand value of the parent firm, or on local reputation (Mathewson and Winter 1985). Finally residual loss may include quasi-rent appropriation (Brickley and Dark 1987, Carney and Gedajlovic 1991, Michael 2000). In this case, franchisees may own assets (as part of the business) that hold higher value with alternative uses, and use this as leverage in contract renegotiation. Alternatively franchisees may demand lower initial fees or royalty rates for higher levels of asset specificity.

Post-Contractual Performance

The majority of franchising research is focused on the ex-ante determination of governance form. However, some more recent work examines the post-contractual operational performance implications of the franchise governance form. The opportunity for the current research lies in the ex-ante incompleteness of contracts and the inability to perfectly monitor ex-post, allowing for residual agency loss. The agency loss component, found to be most prevalent in choosing the franchise outlet form, also serves as the basis for explaining ex-post performance. Existing research on ex-post performance takes on three dimensions; plural form, relational governance, and the fit of strategy to organizational form.

The first dimension of franchise performance research is the limits to adoption of organizational form. In other words, firm performance is observed from the perspective of *plural form*, or one where neither franchised nor firm-owned outlets completely dominate. Research on choice of organizational form describe a plural arrangement, but only as a result of balanced agency factors, transaction costs, or resource limitations. This economic equilibrium view does not recognize that under the plural form, a firm is able to symbiotically exploit unique performance characteristics of both firm-owned and franchised outlets. The plural arrangement permits system-wide uniformity, but also adaptation to local opportunities (Bradach 1997, Yin and Zajac 2004). Through performance benchmarking referred to as ratcheting, both organizational forms benefit from the advantages of the other (Bradach 1997). Managers at firm-owned outlets, motivated by promotion within a corporate hierarchy, are incented to maintain standards and contribute to brand value. Conversely, franchisees have more leeway to try new strategies (Rubin 1978, Bradach 1997, Yin and Zajac 2004, Paik and Choi 2007), and are motivated by residual claims to innovate. More precisely, franchisees are more willing to deviate from corporate guidance if they believe it will contribute to their outlet's profit (Brickley and Dark 1987, Norton 1988a, Bertagnoli 1989, Krueger 1991, Carney and Gedajlovic 1991, Kaufmann and Lafontaine 1994, Michael 2000, Yin and Zajac 2004, de Leeuw, Holweg and Williams 2011). The parent firm, monitoring via integrated information systems, has the capability to benchmark franchisee performance and adapt corporate strategy if franchisee deviations prove effective. Exploiting the advantage of franchisee adaptability can then be achieved by fiat at corporate owned outlets.

If, however, the parent firm observes superior performance at corporate owned outlets relative to franchised outlets, it has limited power to coerce franchised outlets to adopt the effective policy. Contracts permit greater freedom to franchisees, and monitoring is relatively more expensive for franchisees. This introduces the second dimension of post-contractual performance, *relational governance*. Marketing channel theorists posit that non-coercive interactions are more effective at influencing behavior with corporate partners that parent firms have limited contractual influence over (Paik and Choi 2007, Cochet et al. 2008). Bradach (1997) supports the notion that persuasion is far more effective than threat of contract termination or even monitoring for franchised units. High performance of corporate outlets can be a persuasive tool to convince franchisees to similarly uphold standards.

The third dimension of post-contractual performance is *fit of form and strategy*. Contingency theorists observe that the choice of plural form exists as a result of matching compatible strategies and outlet forms (Bradach 1997, Yin and Zajac 2004, Barthélemy 2008). Outlets have agency costs that are more significant to each form, and advantages that are specific to their form. Performance depends on developing a strategy that minimized the form specific residual loss and maximizes the advantages of the plural form.

Hypotheses

The literature predicting governance form informs the extension to the effect of governance form on operational performance. By determining sources of residual agency loss, we can suggest remedies tailored for the conditions of incomplete control as in the

franchise contract model. In this way, we will contribute to research on post contract performance.

Research on the agency theoretical drivers of governance form and more recent work on performance implications of franchising indicate that franchisees, as residual claimants of their local business unit have incentives that focus on local profit rather than the goals of the parent firm and third party distributors, and as a result are more likely to act independently of the franchisor (Brickley and Dark 1987, Norton 1988a, Bertagnoli 1989, Krueger 1991, Carney and Gedajlovic 1991, Kaufmann and Lafontaine 1994, Michael 2000, Yin and Zajac 2004, de Leeuw, Holweg and Williams 2011). Despite the limited previous research on the performance outcomes of franchising focus on aggregated external business measures and not internal indicators, this notion is supported by discussions with the 4PL firm and 3PL DC operator. Therefore, we hypothesize that franchised outlet owners are more likely to order supplies in quantities different from the forecasted replenishment.

> *H1: Replenishment forecast deviation from proposed order levels by a restaurant will be relatively higher among restaurants owned by a franchisee than those corporately owned.*

Similarly, franchising researchers have indicated that franchisees are motivated by profit and not revenue like their corporate manager counterparts. This gives them incentive not to shirk, more actively monitor costs, and waste less (Rubin 1978, Norton 1988a, Norton 1988b, Noren 1990, Krueger 1991). This would imply bias in replenishment forecast adjustments would be negative, and helps the local performance

of a restaurant outlet. However, there are alternative explanations for this hypothesis, and potential for a competing hypothesis.

An alternative explanation for negative bias supported by prior literature is that of incomplete accounting for residual agency costs. Inefficient risk bearing by franchisees could include inventory underinvestment in addition to the more traditional concerns of capital and marketing underinvestment (Brickley and Dark 1987, Carney and Gedajlovic 1991). Free ridership by franchisees could also be an explanation for negative bias in cases where repeat patronage is low or if local competition is not significant. In these cases, negative bias would reflect scarcity rather than efficiency, and hurt local performance. The effect for upstream distributors would be the same, however, so we include this as merely an alternative explanation for our existing hypothesis rather than a separate competing hypothesis. If a franchise owner observes a centrally developed forecast higher than their own perceived knowledge of projected demand, they are more likely than their corporate manager counterpart to revise their replenishment order downward.

> *H2: Bias in replenishment forecast deviations will be relatively more negative among restaurants owned by a franchisee than those corporately owned.*

A competing hypothesis arose from interviews with the focal restaurant firm's promotions team as well as from one 3PL distribution team. They indicate positive expected bias in replenishment order deviations driven by a desire to minimize lost sales by franchise owners. Prior literature could provide support for this view, as particularly in cases where local competition is high (Cochet et al. 2008) and repeat patronage is likely (Norton 1988b), service level and market share considerations could increase

positive replenishment order deviation bias among franchisees to the point of dominating the negative effects described above. We do not measure levels of competition or travel intensity (indicating degree of free ridership) in this work, so do not include this as a formal hypothesis. This could, however, serve as a direction for future research if *H2* is not supported.

Model Development

To test these hypotheses, we built models to predict replenishment forecast deviation and bias with the primary predictor being restaurant outlet *governance form* (GOV), defined and measured as the binary existence of a franchise contract at an outlet (0 indicates a corporate restaurant and 1 indicates a franchisee owned restaurant). Franchised locations are treated as equal, even if a single franchise owner operated several restaurants.

Data Collection and Analysis

We began collecting data with the aim of gathering a sample representative of the almost 15,000 distributed outlets and 8,100 stock units. In the population, there was the potential for more than 120 million daily transactions to evaluate. Data also were stored in multiple segregated repositories, and included a mixture of numeric and non-numeric data. With this combination of volume, variety and velocity, our data were consistent with the most widely used definition for "big data" (McAfee and Brynjolfsson 2012, Megahed and Jones-Farmer 2013, Kitchin and McArdle 2016).

While the "bigness" of data provided tremendous opportunities to explore unexpected relationships, it presented unique challenges and demanded a different

analytical approach. Aside from technical and computational limitations related to the scale of data, described separately in a working paper (Smyth et al. 2017), there are two interrelated inferential challenges in big data analysis. First, the increased size and complexity of big data drive exponential increases in the prevalence of false positives (Waller and Fawcett 2013a). That is, both unsupervised and supervised machine learning algorithms identify (typically) correlative relationships that may have no practical or theoretical meaning (known as epiphenomenality), may be confounded by other unexplored factors (spuriousness), or if practically related, may have a causal implication that is the exact reverse of reality (Darlington and Hayes 1990). The second challenge arises when researchers treat big data as any other, and pursue traditional hypothesis testing and analytic methods. Though theoretically grounded, the results will almost certainly produce a so-called statistically significant result, as the statistical power of a test increases with the size of data (Hair et al. 2006, Wooldridge 2015).

For these reasons, both Waller and Fawcett (2013b) and Cotteleer and Wan (2016) suggest pairing theoretic grounding with big data exploration. A-priori theorizing limits false positives in exploration, and exploration lends credibility to an observed effect by providing necessary context. As such, we explored the additional relationships in our data sample to augment the hypothesis testing of the effect of governance form on replenishment forecast deviation and bias.

To accomplish this task, we required data that was not just "big", but also multidimensional. Only under this condition do complex and unexpected patterns arise. As a result, our goal was to bring in enough relevant explanatory variables so we could

train a process of exploratory knowledge extraction on a representative subset, which could then be applied in cases of larger data in an automated fashion. Data training is a common approach when developing a supervised machine learning algorithm for labeled data (Megahed and Jones-Farmer 2013), and is consistent with the Waller and Fawcett's (2013b) call to combine big data methods with theoretically grounded empirical research.

Variable Evaluation and Selection

Our outcome variables of interest, replenishment forecast deviation ($DEV \in \mathbb{Z}^{+}$) and bias ($BIAS \in \mathbb{Z}$), were both derived from individual replenishment transaction data furnished by the 4PL. Replenishment orders are in units of cases of menu item ingredients. Deviation is the absolute difference between recommended and ordered amounts, and bias is simply the difference. Replenishment forecast deviation and bias was internally tracked by stock keeping unit, individual restaurant outlet, and daily order. Restaurant outlet-level replenishment forecasts and actual ordered quantity were available from a single database.

In an effort to relate this research to the limited extant research on post-contractual performance in firms with franchised or plural form, we also strove to identify common predictors and control variables from extant literature. However, due to the nature of the data in existing research, few commonalities emerged. Previous work either relied on perceptual measures, did not measure *governance form*, were at a higher level of aggregation, or measured information unavailable to us such as multiunit ownership among franchisees (Bradach 1997, Paik and Choi 2007, Barthélemy 2008). Only Yin and Zajac (2004) utilize *governance form* as a variable. Theirs is also the only

existing study that includes data that is directly measured, rather than perception-based. As a result of this limited comparability, we derive control variables and later exploratory constructs primarily from within the available temporal, geographic, and product-based structure characteristics of the available transaction data. In their review of the supply chain forecasting literature, Syntetos et al. (2016) note these grouping dimensions have been shown to effectively account for demand variance, and multiple papers have utilized each to find significant effects (Mentzer and Cox 1984, Fliedner and Lawrence 1995, Zhao et al. 2002b, Zotteri et al. 2005, Williams and Waller 2011b, Rostami-Tabar et al. 2013, Jin et al. 2015, Moon 2015, Paul et al. 2015).

As each deviation represents some edit on the transaction between an individual restaurant and their servicing DC, we included an indicator for which DC would fulfill each order (and would be affected by each deviation). Restaurants are geographically nested in a DC, in that each DC services all stores within a defined geographic area for all products. Advertising is coordinated through a geographically organized group of restaurants, called a cooperative, and since advertising patterns could have a great effect on the variance of replenishment forecasts (and resulting deviations) for a product, we need to track the advertising cooperative to which each store belongs. Stores are nested in cooperatives much as they are in DCs, but DCs and advertising cooperatives are not nested in or coincidental to each other. That is, a cooperative of restaurant outlets could be serviced by multiple DCs, and vice versa. Cooperative membership is geographically nested in a television market, which is then geographically nested in a region. Therefore

we included cooperative membership, television market, and region for each transaction from a separate database.

As our measures of interest were scale dependent, we needed to include controls for restaurant throughput. Historical usage ($HIST \in \mathbb{R}^{+}$) was included as a scaler for the level of consumption at a restaurant outlet over the previous replenishment period. Restaurants with high historical usage tended to have higher deviation levels, but merely as a function of higher throughput volume. The same is true for recommended order quantity for the upcoming replenishment period (known as proposal amount or $PROP, \in \mathbb{R}^{+}$), though in a manner somewhat independent of historical usage. For this reason, we included both *PROP* and *HIST*.

Training Data Set Formation

As described in Megahed and Jones-Farmer (2013), model building in an environment with large multidimensional data can quickly become overwhelmingly complex. That is why they recommend training a model on data that is relatively simple, yet can be expected to represent effects beyond the limited scope. When gathering a sample to train a big data exploratory model, we had to strike a balance between complete representation and tractability. If our sample was too large, we may get bogged down by limitations relating to computational complexity. If our sample was too small, we run the risk of identifying anomalous effects and limiting the general applicability of our results. To achieve this balance, we limited the stock keeping units, date range, and restaurants examined to be a representative sample of a wide range of menu offerings, demand patterns, geographic regions, and seasons. The representativeness of our sample

was achieved through multiple discussions with the 4PL firm's analysts. In this way, we could develop a machine-operated process guided by domain knowledge. We could test theoretically-based hypotheses on a limited range of data that can then be applied via a supervised machine learning algorithm to the exhaustive set of transactions for use by the restaurant firm and 4PL.

We included one full year to capture the seasonal effects and annual promotions observable in the data. We also limited our sample to one year of data, as firm forecasters were still in the process of refining their approach to restaurant replenishment forecasts, and the process had changed in the previous two years. As a result, the types of information collected before the identified range differed in nature and quality. Following this period, increased competitive pressure caused the parent firm to make significant menu changes. Therefore, data after the identified range differed in products offered and in maturity of product demand. We selected 39 of over 8,100 stock keeping units, representing multiple of the highest volume limited promotional items, perishable refrigerated items, frozen items, fresh produce, meats, dry goods, condiments and paper products. The restaurant firm's 4PL indicated that this sample was representative of all major demand patterns they experience in their forecasting. Similarly, we selected the restaurants serviced by only 8 of 34 domestic DCs. This resulted in only 4,173 restaurants spread fairly evenly across the contiguous U.S., with the greatest concentrations of outlets being in the greater Los Angeles, Chicago, Washington D.C. and Dallas-Fort Worth metropolitan areas. Again, the firm's 4PL partner indicated this was a representative sample of DC locations and operators. The result was a more

manageable sample of 15.5 million observations, each representing an individual transaction between a restaurant outlet and their servicing DC. Training samples would then be randomly drawn from this pool, depending on the complexity requirements of the analysis method, and retaining some portion of the data as a holdout or verification sample.

Data Processing

As noted in Cotteleer and Wan (2016), rich datasets from a large company with unintegrated databases can pose significant difficulties and require a tremendous amount of time to process and understand, especially in the early stages of analysis. Understanding the data meant understanding the business operations that drive data generation; by whom, how, and for what reason each element of data was recorded. It also meant reconciling inconsistent labeling, definitions and usages for data elements between the various repositories we drew from. For this we are grateful to the steadfast (and patient) support from the 4PL analysts and corporate liaison as we gained an understanding of the vast pool of data we were analyzing. Their assistance was integral as we assessed data quality, integrated data samples from each separate source, recoded and parsed our data prior to our exploratory analysis.

Data entry errors, database anomalies and other intrinsic quality issues are magnified in big data. As manual scrubbing of data for errors is not possible, we calculated descriptive statistics for all numeric indicators, as well as factor summaries for all non-numeric indicators to identify anomalies. We relied on the guidance of the 4PL's analysts to determine what constituted an "unreasonable" value in order to assess whether

a mathematical outlier must be cleansed from the data. If any null cells or unreasonable values existed, the value was recoded (if evident the null indicates zero), or the observation was eliminated (if an unexplainable value).

Integrating or merging data may also be more difficult with big data, as information gathered for different purposes even within the same business may have differing levels of aggregation, dissimilar time windows and date coding, and may have no common features required for seamless merging. We merged our data by identifying common factors in the multiple databases we drew samples from, making sure along the way that definitions of these common factors were consistent. Often this entailed incorporating data from unrelated repositories solely as a "cypher" linking data elements we *were* interested in. This also required frequent recoding of date formats, stock unit identifiers, and location codes in nested levels.

The process of merging multiple databases left many redundancies and irrelevant indicators, so parsing was a necessary step. The remaining variables after parsing were a replenishment forecast deviation and bias (*DEV, BIAS*) terms as the dependent variables, governance form (*GOV*) indicator as the primary independent variable, the scaling variables *HIST* and *PROP*, multiple nested multicategorical location indicators, date indicator and multicategorical product indicators.

To numerically analyze multicategorical variables, they had to be transformed into indicator variables for each category level. This increased the number of independent indicators to be roughly the sum total of factor levels in each of the original variables. We chose to use a weighted effects coding scheme, as each treatment

combination did not have equal sample sizes. In this scheme, the numeric value of the indicator variable represents the deviation from the overall mean through membership in a level of a multicategorical variable. Any multicategorical variable with g levels becomes $g-1$ indicator variables (D_1 through D_{g-1}), where levels 1 through $g-1$ are denoted by only one indicator variable taking the value of 1, and the rest 0. The reference level g is denoted by all $g-1$ indicator variables taking the value of a counterweight $-h_j/h_g, \forall j \in \{1,2 \ldots, g-1\}$, where h_j is the number of observations in each treatment level 1 through $g-1$ and h_g is the number of observations in each treatment level g (Darlington and Hayes 1990, Cohen et al. 2013).

Data Exploration

After data processing was complete, we could begin exploring the sample for contextual factors, clusters, or variables to better characterize the effect of governance form on deviation and bias. Our collected data permitted exploration over a limited range of products, restaurant outlets and dates, such that a more general pattern could emerge. We attempted multiple extant data mining approaches as suggested in Hand et al. (2001), Han et al. (2011), and Kuhn and Johnson (2013), with the intent of demonstrating effective data grouping and reductive techniques scalable to the population, and replicable in different contexts. These group constructs would then be included, along with governance form indicators, in a regression model predicting forecast deviation and bias. We had limited success with these approaches due primarily to the scale of our dataset. The exploratory approaches we tried, and ultimately abandoned, are described in more detail in a separate working paper (Smyth et al. 2017). A brief summary of our

efforts is listed below. The primary subdivisions among the exploratory approaches are observation-based and covariance structure-based grouping mechanisms.

We employed the observation-based grouping mechanisms of k-means and both agglomerative and divisive hierarchical clustering, but were unsuccessful for two reasons. First, a distance matrix must be calculated between all *n* observations, thereby increasing the temporary or random access storage requirements. For samples as small as one million, storage requirements are on the order of terabytes or more (Buchholtz 1962); well beyond the current capabilities of most computers. This meant that our training sample size was limited in size. The second and even more hindering reality is that observation-based methods require manual interpretation of a stable structure. When attempted, the resultant clusters were neither stable nor interpretable, so we moved on to covariance structure-based grouping mechanisms.

Factor analysis, our chosen covariance structure-based approach, had the computational advantage of requiring less temporary memory (Sharma and Paliwal 2007) than observation-based methods. This is because they identify structures that may exist between *m* predictor variables, typically far fewer in number than *n* observations. This method also has the advantage of established statistical indicators to aid selection of factor models (Horn 1965, Velicer 1976, Zwick and Velicer 1986). Unfortunately as with observation-based methods, structures were neither stable nor interpretable, and each statistical indicator specified a different model.

The instability and uninterpretability of models developed by both observation-based and covariance structure-based exploratory methods was likely due to complex and

unexplainable interactions of effects due to high dimensionality (Gu et al. 2012). In response, we shifted our exploration to be based on a more parsimonious set of temporal, geographic, and product-based structure characteristics in regression-based exploratory modelling.

A Novel Regression-Based Approach

Making use of the hierarchical structure of the available data, we formulated a process to predict replenishment forecast deviation and bias from governance form, as well as temporal, geographic and product indicators. In this method of analysis, we only estimated first order linear effects. While this omits the more complex interactions that may exist between variables, it also greatly reduces the manual interpretation of the vast number of variable combinations that are possible in higher order effects. By eschewing interaction effects, we avoid one of the major pitfalls of massive and highly dimensional data (Gu et al. 2012), and permit additional scalability of our model (National Research Council 2013). In terms of computational parsimony, regression models estimated by ordinary least squares (OLS) are highly scalable (providing that $m \ll n$, and Cohen et al. 2013 recommend $n/m \geq 40$ for data-driven hierarchical analysis) when compared to interdependence techniques such as cluster and factor analysis.

The concept behind the approach is that initially only aggregate level nested dummy indicators for categorical variables such as region or month are used as predictor variables in linear models with replenishment forecast deviation and bias as the dependent variables. Only the indicators with the highest aggregation are initially included to limit the number of terms to evaluate, similar to the *fixed effects approach to*

clustering described in Cohen et al. (2013). Should a term prove to be a significant predictor of deviation or bias, we remove only the significant predictor and replace it with the nested disaggregated indicators contained in it, for instance television market in place of region and week rather than month. As the data are nested, and we are already using a weighted effects coding scheme, we can interpret the other unremoved coefficients as nearly equivalent to their interpretation in the original model with only aggregated terms. This iterative process requires an evaluation of terms after each stage of disaggregation, with the number of iterations being dependent on the number of hierarchical tiers present in the data. As terms are iteratively disaggregated, we approach a model similar to Cohen et al.'s (2013), *disaggregated analysis* that has completely atomized indicators and numeric independent variables. This method, that we call *hierarchical progressive disaggregation (HPD)*, permits an analyst or manager to identify relatively few aggregated identifiers that have some significant effect prior to diving deeper. This is particularly useful as this method is scaled up from our sample with only 39 disaggregated product indicators and 4,173 disaggregated restaurant outlet indicators to one that potentially includes all 8,100 stock units, nearly 15,000 restaurant locations, and longer time windows.

We estimated separate models for replenishment forecast deviation (*DEV*) and replenishment forecast bias (*BIAS*). To estimate these two dependent variables, we initially fit linear regression models via OLS. Under this estimation method, we had to satisfy certain assumptions: (1) linearity in the relationship between independent and

dependent variables, and in the residuals, (2) normality, (3) homoscedasticity and (4) independence (Cohen et al. 2013).

Predictors in our model that are multicategorical were subsequently recoded as weighted effects coded binary indicator variables, so the assumption of linearity for these were assured. With only two treatment levels possible for each condition, we would not be able to characterize a more complex relationship between independent and dependent variables. However, *HIST* and *PROP* both are either continuous or have enough treatment levels where nonlinearity could be an issue. Both *DEV* and *BIAS* are counts, but deviation is strictly non-negative whereas bias can be either negative or positive. Both counts also include a large proportion of observations with zero values, meaning the restaurant accepted the centrally developed replenishment forecast. Of note, we had eliminated the multicategorical DC indicator during previous exploratory analysis due to high multicollinearity with other location indicators. This made factor models inestimable. The terms were also left out of subsequent analyses both for high collinearity and because DC indicators were not nested like location indicators restaurant, cooperative, and region.

When fitting initial OLS models for deviation and bias, we found that our residuals exhibited non-normality and heteroscedasticity. While both sets of residuals still had the characteristic bell shape indicating normality, they were bimodal with a heavy concentration near zero. This indicated error in the model prediction when the true deviation or bias was zero, with a more normally distributed second error source. Plotting residuals against predicted values, we found similar tight groupings around zero,

with a more homoscedastic grouping of observations away from zero. This result implied model misspecification, and as a result we had to explore alternative model conceptions.

Cohen et al. (2013) recommend generalized linear models (GLM) to account for such heteroscedasticity, and as deviation occurs as positive integer counts, we initially pursued a Poisson distribution for the outcome variable. We encountered two main issues with fitting such a model. One is that the computational complexity in iteratively fitting a ML regression equation exceeds that of OLS (Minka 2003) and lacks scalability (Toulis and Airoldi 2015). Even with a (relatively) small training sample of five million, the difference in computation time between ML (using Fisher scoring or Newton-Raphson iteratively weighted least squares algorithms) and OLS (using the QR decomposition algorithm) is quite noticeable (Fox and Weisberg 2010). This of course would be exacerbated in the larger enterprise-sized samples, though can be mitigated partially by capping the number of iterations in the ML estimation (Cohen et al. 2013). The second issue is that a Poisson model ended fitting poorly to our data. Standard indicators of fit will tend to fail (asymptotically) as sample size increases (Maydeu-Olivares and Garcia-Forero 2010). Fit indices are based on the assumption of an approximate fit to a theoretic distribution. As sample size increases, the credible interval window will narrow and a null hypothesis will invariably be rejected, thus violating the assumptions of such a model. Observed phenomena rarely (if ever) converge perfectly to a theoretic shape, and as the fidelity of the sample grows, the likelihood that it will deviate from a theoretic distribution shaped (in this case) by only one parameter increases. It is likely that a Poisson distribution poorly approximates a non-negative

observed value that is actually an amplified reflection of negative and positive count values.

Jöreskog (2002) notes that this particular data structure may better be approximated with a Tobit model that assumes a censored normal distribution in the response variable. A histogram of deviation in our data appeared to support this assumption, with a large proportion of values (~70%) accumulated around the censored tail at zero. Unfortunately, this type of model shares the same weaknesses of a Poisson model in that it is estimated by ML, so lacks scalability (Jöreskog 2002) and asymptotically violates its (in this case three) parameter assumptions (Henningsen 2010).

Accounting for Cases of Non-Deviation

In our attempts to characterize the data via ML models with parametric assumptions, we began to recognize a different character in those orders in which no deviation or bias occurs. As mentioned previously, heteroscedasticity occurs mainly around the zero values, indicating that linear or parametric models estimate non-zero values fairly well, but that zero values are potentially determined by a separate function. This problem is known as nonrandom sample selection driven by incidental truncation (Wooldridge 2015). If this truncation is ignored and we attempt to estimate a parametric model of the whole sample, there exists a sample selection bias (in a direction determined by the nature of truncation). One method for correcting this bias is via the Heckit method, which estimates the existence and magnitude of a dependent variable in separate equations. The existence estimate is accomplished via logit or probit regression on the whole sample, which permits estimation of an inverse Mills ratio $\lambda(z_i\gamma_i), i \in \{1{:}\, m\}$. In

this function, z_i represents the list of variables predicting s, the existence of deviation (which contain x_j, the predictor variables for y, nonzero deviation magnitude), γ_i is the list of regression coefficients. The magnitude model is then estimated via OLS with the Mills ratio included as $E(y|z, s = 1) = x_j\beta_j + \rho\lambda(z_i\gamma_i), j \in \{1: k \in m\}$, and ρ is the correlation between error terms of the two models (Wooldridge 2010). Models estimated in this way have been found to be both consistent and unbiased (Heckman 1976).

This sample selection observation has a rationale beyond data mining and exploration of residuals. A restaurant outlet manager or franchisee may be predisposed to avoid deviating from the forecasted replenishment under certain conditions that are independent of the conditions that affect how *much* they would deviate. Perhaps in instances where they have low volume in or low proportion of profit derived from a particular product, or little past knowledge of the sales performance of an item, they would defer to the centrally developed replenishment. Consequently, we would not expect these (or other) factors to have the same effect for an observation in which no deviation or bias occurs as we would for those in which we observe at least some deviation or bias. With such a large proportion of observations with no change to the proposed forecast, it may make more sense to truncate those observations rather than censor or try to parameterize them in a single model.

We coded a new variable to be a deviation existence indicator ($DEVEX \in \{0,1\}$), and estimated a logit model to try and predict *DEVEX*. Utilizing the *purposeful* method of fitting logit models (Hosmer Jr. et al. 2013), our preliminary main effects model retained all predictor variables included in the original OLS model. This is based on their

Wald test significance and consistent with their logical and theoretic justification for inclusion. Though removing terms with the lowest Wald test p-values did not heavily influence the remaining regression coefficients (change of $< 15\%$), nested models were significantly different from each other as measured by the likelihood ratio test. Therefore we retained all original predictors. Evidence of a main effects model was supported by estimating a model via the method of fractional polynomials, which indicated that deviance is minimized when only first order terms are included. Thus the preliminary final model contained only predictor variables included in the original OLS model, and the model's log-likelihood indicates a significant increase in deviance explained over a null model ($p \ll 0.001$).

The logit model indicated franchised outlets were 39% more likely than corporate restaurants to deviate from a forecast, and that this effect was highly significant ($p \ll 0.001$). This is further support of cross-tabulations and descriptive statistics that indicate increased likelihood among franchisees. The mosaic plot in Figure 2 indicates a larger proportion of orders placed by franchisees were non-zero when compared to orders placed by corporate managed outlets. The width of the mosaic tiles serve to indicate the relative number of franchised and corporate transactions. 89.5% of observations in our sample were from franchised restaurants (which made up 87.5% of sampled locations).

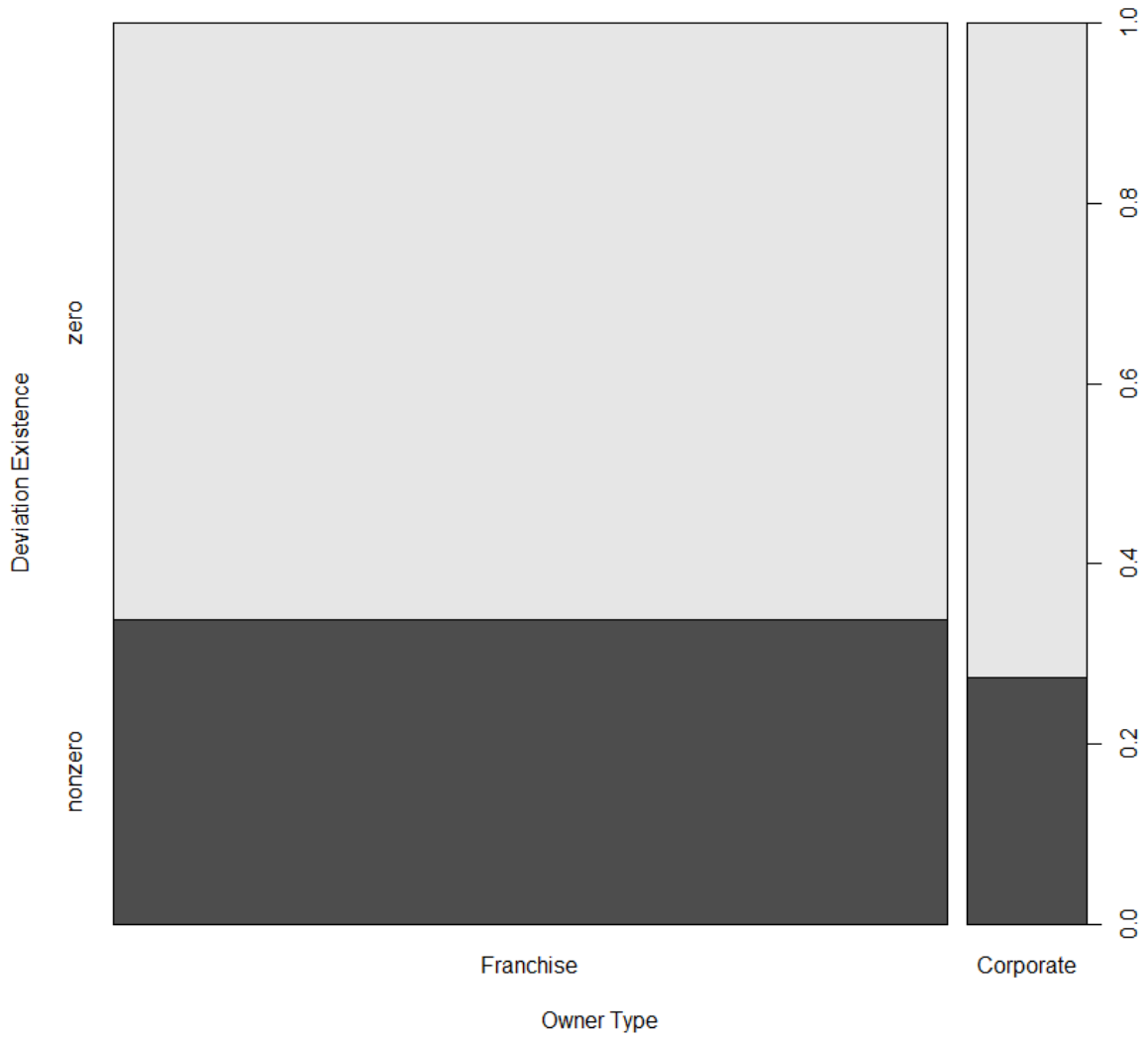

Figure 2: Prevalence of Deviation Existence by Governance Form

However, our logit model had marginal discrimination ($AUC = 0.649$), meaning that given two observations where one has zero deviation and the other has non-zero deviation, the model will only correctly classify these observations 65% of the time. Beyond classification weaknesses, our model also had a number of indicators of poor model calibration. LOWESS smoothed plots of numeric (treated as continuous) variables *HIST* and *PROP* against *DEVEX* reveal monotonic behavior, which would be expected in

a properly fitting logit model. However, when logit transformed, these plots exhibit nonlinearity in the lower range of values, indicating possible model misspecification. Both the Pearson's Chi-Square and the Hosmer-Lemeshow goodness of fit tests indicate poor model fit ($p \ll 0.001$). Similarly, the McFadden pseudo R^2 measure that represents a proportional reduction in error variance (not to be confused with the R^2 measure from OLS) indicates sub-par explanatory power of the model at $R^2_{McF} = 0.048$ (Wooldridge 2010). Due to these limitations, we cannot have a reasonable estimate for the Mills ratio. This lack of fit may be due to the lack of an exogenous variable in $z_i \notin x_j$ (Wooldridge 2015), and requires the assumption (that we later examine) that $\rho = 0$.

We then conducted a separate OLS analysis only on those observations in which a deviation occurs. We did so following the *HPD* procedure described previously. The initial model is the most highly aggregated, with indicators only for month ($MONTH_c, C = \{1,2, \dots 12\}$), region ($REG_k$) and an aggregate indicator for product grouping (GRP_r). Every day ($DAY_a, A = \{1,2, \dots 365\}$) is nested in a week ($WK_b, B = \{1,2, \dots 52\}$), that is then nested in a month by the month a week began. All $I = 4{,}173$ restaurants ($REST_i$) are nested in $J = 56$ advertising cooperatives ($COOP_j$), which in turn are nested in $K = 16$ geographic regions. Each of $P = 39$ products ($PROD_p$) fall into $R = 6$ aggregate product grouping indicators that we jointly determined with the restaurant's 4PL to be most likely to have common advertising, storage and handling characteristics. Frozen, refrigerated, paper, dry goods, promotional items and bread products were all determined to be unique groupings that would be expected to behave similarly in terms of replenishment forecast deviation and bias. By initially using such

aggregated terms and estimating only first order effects, we were able to initially evaluate and interpret only 34 regression coefficients. The initial evaluated model for deviation is of the form:

$$DEV_{aip} = \beta_0 + \beta_1 GOV + \beta_2 HIST + \beta_3 PROP + \beta_{3+c} MONTH + \beta_{3+C+k} REG + \beta_{3+C+K+r} GRP$$

And the model for bias is simply:

$$BIAS_{aip} = \beta_0 + \beta_1 GOV + \beta_2 HIST + \beta_3 PROP + \beta_{3+c} MONTH + \beta_{3+C+k} REG + \beta_{3+C+K+r} GRP$$

Outlier Identification

As with the OLS models that included *all* observations, we estimated separate models for bias and deviation, but now among transactions with strictly *nonzero* values. While fitting the models, we tested for and observed outliers in both models based on the global influence measure Cook's Distance $D_i = \frac{\sum(\hat{Y}-\hat{Y}_i)^2}{(m+1)MS_{res}}$ with m predictors (Cohen et al. 2013). We selected a global influence measure over a specific influence measure such as DFBETA because estimation requires *m* times less computation time in the global measure. This was significant given our training sample size of five million. While many cutoff thresholds are proposed as being worth examining, we selected a value $4/(n-m-1)$ with n observations and m predictors, which is more appropriate for large data samples (Fox and Weisberg 2010). Despite this being a more stringent cutoff designed to limit the amount of outliers an analyst must examine, the deviation and bias models had 53,144 and 58,023 respectively (out of 1.6 million observations). In the end, four observations were removed based on Cook's distances that were several times larger than all other highly influential observations. While it is unwise to remove observations

simply due to their large influence (Hair et al. 2006), we observed these four transactions also to have order values 23.5-91.7 times larger than historic usage and 76.3-80.2 time larger than the proposed order size. These are not reasonable values and constitute obvious entry errors. Worth noting: while outliers manifested in a fairly proportional manner to all predictor factor levels, two stood out. Bun and fry transactions made up 91.8% of the most influential observations in the deviation model and 93.2% of the most influential observations in the bias model. In fact, fully 40.3% of bun and 15.7% of fry transactions in the deviation model and 43.7% of bun and 18.5% of fry transactions in the bias model were considered "outliers". This indicates that these products behave differently than the others in this model, and that the linear prediction may not be sufficient for these products.

Results

After removing erroneous outliers, we began the disaggregation process and report the results starting with the deviation model. Starting with terms at the highest level of aggregation, we fit a model that explains 50.3% of the variation in order deviation. The results of the aggregated model indicates nearly all predictors had a significant effect by traditional alpha-level cutoff standards. This is to be expected, as our sample size is extremely large. In fact, as this process is scaled to larger portions of an enterprise dataset, it would be unlikely to calculate an effect that wasn't significant by traditional statistical cutoff levels (i.e.: $\alpha = 0.05, 0.01$ or 0.001). Using any fixed level of statistical significance as a threshold becomes problematic in large data sets, because significance depends not only on effect size and dispersion, but increasingly on the

number of observations in any treatment level (Hair et al. 2006, Wooldridge 2015). We therefore used a relative threshold to select and disaggregate only those factors whose effects are least likely to be due only to chance. As governance form is the focal predictor in this model, we defined statistical significance of covariates in relation to the statistical significance of that term. Alternatively, if the analysis is purely exploratory, or if there is a much larger set of variables, some percent of the most statistically significant terms may be candidates for disaggregation.

Deviation Model

For the deviation model (when deviation is nonzero), only two regions and two product categories had higher significance than the effect of governance form. For the first disaggregation, we replaced the indicators for the Heartland and Southern California regions and the frozen and refrigerated product categories with the 16 and 15 respective nested terms that make up those aggregated categories. We then re-estimated the model with the disaggregated terms. Not surprisingly, the disaggregated model explained more of the variance in deviation at 52.0%. In this model, deviation in orders made by franchised outlets are expected to be 8% higher than those made by corporate outlets, all else equal. We should expect a greater number of terms to represent a higher proportion of variance given that all information contained in the aggregated terms is implicitly contained in the disaggregated replacement terms. By again observing only some subset of terms (bound by either the significance of a focal term or some percent of variables), we see the value of such a hierarchical procedure. Restaurants orders were on average

3.1 cases different from recommendations when they made replenishment forecast revisions, so all effects are cited relative to this.

The partially disaggregated location indicator narrowed the search for the source of variance from two very expansive regions, to many of the constituent marketing cooperatives. The result is that the Los Angeles cooperative is now the only location whose effect is less likely than governance form to be simply attributable to chance. Orders in the Los Angeles cooperative deviate on average 0.3 cases less than other cooperatives and regions, all else equal.

Interesting results also come from disaggregating product terms. The disaggregated model includes 3 product groups and 20 individual products that have a greater significance than governance form. This result is slightly more complicated in its interpretation, as three indicators (representing the dry goods, bun and paper product categories) that had previously been less significant than franchising are now more significant with the removal of some terms and the addition of others. This relative change is due to collinearity that exists between removed and retained variables. As indicators of product category that were removed are mutually exclusive of product category indicators that remain, it is only the removed indicators of region that covaried with the remaining product category terms. In the same vein, the additional cooperative indicators (as nested product indicators also are mutually exclusive of product category indicators) covaried less than the original aggregated terms and so had less of a confounding effect on the remaining product category indicators. This result provides useful information about the effect of specific product categories and products on

replenishment forecast deviation. For instance, among those significant predictors (relative to governance form) only small beef patties, buns, fries and ice cream had a positive effect on deviation (0.9, 2.7, 3.7 and 0.5 cases respectively). However, these items constituted 69.2% of all products ordered at the restaurant level. In essence, the products that have the greatest volume of orders (and thus likely the greatest impact on costs) are driving positive deviation by the greatest extent.

As we are observing independent variable significance relative to the significance of governance form, we should also note changes in the significance of governance form as a predictor of replenishment forecast deviation. The likelihood that the effect of governance form on deviation is likely to be due only to chance has increased with the introduction of disaggregated indicators. This is because the newly introduced indicators covary with governance form. We should expect such confounding with the introduction of additional terms. If governance form is heavily confounded by the introduction of additional terms, then a more stable point of comparison may be desirable (such as some fraction of the most significant terms, as suggested above). For our current analysis, however, we retain a consistent logic for selecting terms to disaggregate.

Bias Model

For the model predicting nonzero bias, two of 12 months, eight of 16 regions and five of six product categories had a significant effect on bias relative to the significance of franchising. The aggregate model explains 46.1% of variance in nonzero bias. After disaggregation, 54.8% of variance is accounted for by the model. In this model, bias in orders made by franchised outlets were 3% higher than those made by corporate outlets,

all else equal. This increase in explained variance is expected when expanding the number of temporal, geographic and product-based terms via disaggregation. Restaurants tended to revise their orders up by an average of 2.2 cases if they revised a replenishment forecast, so all effects are cited relative to this.

The partially disaggregated time indicator provided additional fidelity in determining which time periods may have significant effects on order bias. Five months that had previously had less statistical significance than governance form, as well as seven weekly indicators were now relatively more significant in predicting order bias. This increase of significant terms had less to do with the disaggregated covariates than it had to do with the decreased significance of franchising. After disaggregation, it became 8.043×10^{31} times more likely (though still with $p \ll 0.001$) that the effect of governance form on bias was purely due to chance. The introduction of disaggregated terms confounded governance form's effect to a greater extent than the retained aggregated terms, thus reducing its apparent effect. Additionally, the removal of aggregate non-temporal terms that acted as confounders caused the retained aggregate time predictors to gain significance. The greatest effects occurred in the second week of May (0.5 cases larger), and three of four weeks in December, all of which coincide with a significant negative bias effect (0.4, 0.6, and 0.7 cases smaller respectively). This disaggregation has identified multiple more specific time periods to evaluate possible reasons for the observed effects. For instance, the advent of the holiday season may signal lower levels of sales throughout the system that are not currently being captured in the restaurant firm's demand planning process. Individual restaurants are recognizing

this, but for some reason the corporate forecasters may not. Worth noting, we do not compare the bias in restaurant level orders to actual sales data, so it may be some cognitive dissonance in the individual restaurant owners and managers (and not in corporate demand planning) that is driving this observed behavior.

Location indicators in the partially disaggregated model included 3 regions that were not previously significant, as well as 25 advertising cooperatives as statistically significant. Among those regions and cooperatives considered statistically significant, the largest positive bias effect occurred in two cooperatives between Oklahoma City and Dallas (1.7 and 1.2 cases higher on average respectively), as well as two in central Missouri (1.4 and 1.5 cases). The largest negative bias effect was experienced in a cooperative in rural west Texas (1.2 cases), but also the two urban cooperatives in Las Vegas, Nevada and Los Angeles, California (0.7 and 0.6 cases).

Finally, in the disaggregation of product terms in the model predicting bias, one previously non-significant product category and 33 individual products were relatively more significant in predicting order bias than governance form. In fact, the only non-significant product term was lemonade. The largest positive bias effect was observed in buns (2.1 cases) and fries (2.8 cases), whereas the largest negative effects manifested in napkins (1.1 cases) and apple slices (0.9 cases).

Discussion

These results have implications for our hypotheses regarding the effect of governance form on replenishment deviation and bias, extensions to post-contractual

performance management, and methodological implications from our proposed technique of disaggregation.

Effects of Governance Form

Through our analysis, we found mixed support for our hypotheses. *H1* was supported; replenishment forecast deviation from proposed order levels by a restaurant was relatively higher among restaurants owned by a franchisee than those corporately owned. This finding is consistent with both restaurant firm expectations and extant literature on governance form, but now applied to internal measures of supply chain performance (deviation).

However, *H2* was not supported; bias in replenishment forecast deviations was relatively *less* negative among restaurants owned by a franchisee than those corporately owned. This would seem to support the proposed unofficial competing hypothesis, where service level and market share considerations could increase positive replenishment order deviation bias among franchisees to the point of dominating the negative bias effects from improved oversight and frugality, or inventory underinvestment from inefficient risk bearing or free ridership. Further work to confirm this alternative explanation would require additional measurement of proxies for proportion of wealth tied to a franchise such as multi-unit ownership (indicating degree of inefficient risk bearing), and levels of competition or travel intensity (indicating degree of free ridership).

This unexpected result could also be the result of greater levels of unreported local promotions by franchisees, or improperly reported inventory levels. Alternatively, as we did not assess the accuracy of the proposed replenishment forecast against point of

sale consumption, the franchisees could be compensating for a systematic underestimation of demand. It could also be some combination of these causes. Each of these possible alternative explanations challenges the accepted paradigm among governance form scholars that franchisees more parsimoniously steward their resources, and merits further investigation.

Extensions to the Post-Contractual Effects from Governance Form

Firms that utilize franchise governance can use these findings on deviation and bias to address how internal order deviations are structured in contracts. This could be implemented in future contracts as a limit to how much an order can be edited, or a simple cue so that changes over a certain threshold require an explanation or update of previously misreported local inventory or promotional activity. Parent firms can also implement these findings in ways described in recent research on post-contractual franchise performance, as described below.

Those firms that utilize the plural form of governance can use these findings (or rather, employ these methods to their own replenishment data) to ratchet franchisee order edits if it is found that they improve replenishment forecast accuracy relative to point of sale consumption (Bradach 1997, Yin and Zajac 2004). By recognizing local information advantages, parent firms can incorporate the information in an attempt to improve replenishment forecasts, which in turn requires less system wide inventory to account for variance.

Conversely, if it is found that franchisee order edits degrade accuracy relative to point of sale, this provides evidence to convince franchisees to act in their own best

interests. Given that franchise contracts provide for less coercive control by the parent firm, such information is integral to the relational governance proposed by Bradach (1997), Paik and Choi (2007), and Cochet et al. (2008).

Finally, knowing that franchisees exhibit higher deviation and more positive bias can aid the restaurant firm in aligning fit of form and strategy. If the cost of this increased variance in their replenishment system outweighs the marginal benefit of franchising, the restaurant firm may consider changing their policies for managing franchisees or their mix of franchised restaurants. Bradach (1997), Yin and Zajac (2004) and Barthélemy (2008) all suggest properly accounting for and mitigating agency costs particular to a governance form can contribute to minimizing form-specific residual loss and maximizing the advantages of the plural form.

Methodological Implications

The contextualizing effect of *HPD* can help target responses to only those temporal, regional or product category peculiarities that have a distinct positive or negative effect. For instance, knowing that replenishment forecast deviation is much higher among only small beef patties, buns, fries and ice cream or that positive bias is driven overwhelmingly by buns and fries may reduce unintended consequences of broader policy changes. The higher bias observed in mostly rural Midwestern areas may be either systematic under-forecasting, or could be a response by restaurant outlets for unreliable service from a common distributor. This allows the parent firm to target resources to either improve point of sale demand forecasts, or to evaluate potential poor service of a regional distributor. The same is true for identifying benchmarking

opportunities. The lower observed deviation in the Los Angeles cooperative, and lower bias observed in many urban cooperatives can serve as an example for how to structure future contracts, or how to interact with restaurants after contracts are established. Regardless of how these outlet level order edits are affecting point of sale demand accuracy, deviation and bias is causing additional variance for upstream distributors.

The results indicate that *HPD* provides targeted information to managers at the restaurant firm with minimal computation and human interpretation in a supervised learning process. This principle can be extended to any large enterprise with an independently nested structure. This includes most restaurant chains like the focal firm of this study, retail chains, but also firms that provide primarily services rather than goods. Take for instance a large cable company trying to forecast the consumption of cable during its service calls. In terms of regional and temporal aggregation, they may observe higher consumption in northern regions during periods with known severe weather that may damage lines. This would indicate they should provide offices region-wide with greater supplies of cable and possibly shift their staffing. After disaggregating the most highly significant terms, they may find that the variance of the region is actually driven by a single office or small subset of offices. This would indicate a different response, perhaps related to local management or training. While this example is fictional, it demonstrates the possibility for surprising insights through *HPD* for alternate large hierarchical organizations where data volume makes manual and even computational identification difficult. The process is also not limited to those firms that

use the franchise contract model, as any factor to be examined can be contextualized by pairing with such an exploratory process.

In this paper, we demonstrate only one level of disaggregation in our model, though the data structure of our temporal and geographic indicators would permit additional iterations as they both include three nested levels. This is because we wish only to demonstrate the potential value of *HPD* by showing that it can parsimoniously isolate heterogeneous effects. We do not test any theories about regional, temporal, or product-based effects, so end our analysis at one level of disaggregation. Our models predicting deviation and bias in replenishment forecasts are a small-sized example, but it is evident the value the *HPD* process can bring to a large sized data set. Instead of an over-specified model with thousands of confounding indicators, as would be the case if the lowest level of disaggregation were used, the analysis begins with a relatively simple model with coarse indicators. Besides being easier to interpret, this aggregate model is more scalable, given the time complexity of calculating an OLS regression model in a population of billions (or more) of transactions and thousands of indicators. The disaggregation process then can refine a search based on coarse indicators that demonstrate a significant effect on the dependent variable (in our case replenishment forecast deviation or bias). This process can be set up as a simple machine learning process as it uses simple logic to disaggregate. The "supervised" portion of the process is then an evaluation of the various iterations of disaggregation. Each level will be an increasingly complex model to interpret, so it is up to the analyst to decide when to end

the process. The nested structure of the data also acts as a natural termination for the process.

Limitations

This study provides contributions to both management theory and explanatory methods. However, as with all research, it suffers from a number of limitations. The initial limitation is that the sample this analysis is based on comes from a single large quick service restaurant firm. While this limits generalizability of our results, there are several mitigating factors. First, we include data from over 4,000 geographically dispersed restaurant outlets in our model, which accounts for about 1.9% of all domestic quick service restaurant outlets (Census Bureau 2012). Second, we worked with the restaurant firm's demand planners to sample products that represent a wide range of demand, advertising, storage and handling characteristics. Finally, as an industry leader (Hoover's 2016), the characteristics of this firm's outlets can be expected to represent large portions of the industry.

A second limitation is the use of a truncated sample examining only transactions with non-zero values of deviation and bias. This requires the assumption $\rho = 0$, or that the error terms of the two models in the Heckit method are uncorrelated. This assumption holds if model coefficients are observed to be consistent in the truncated sample (Wooldridge 2015). Using a random holdout sample of five million transactions, we observed an OLS model with the same aggregate terms found to be significantly related to changes in deviation and bias. Coefficient estimates were directionally identical between samples and changed by at most a few percent (acceptable under

thresholds established in Hosmer Jr. et al. 2013). This represents a very small effect difference and so can be considered consistent.

This truncation relates to a third limitation of our study. When fitting GLM models, parametric assumptions fail asymptotically as sample size increases (Maydeu-Olivares and Garcia-Forero 2010). As researchers set out to explore larger datasets, traditional fit measures may indicate rejection of model types that are logically and practically appropriate for representing the data. Future research needs to address this weakness in fitting GLM models to larger sample sizes.

A fourth limitation of the study is that it ignores complex interactions and higher order effects. It is likely that some combinations of factors have significant effects on deviation and bias. However, in an effort to limit required manual interpretation of complex interactions as *HPD* is scaled to massive datasets with thousands of indicators, we purposely limit our scope to first order effects.

Finally, deviation and bias in restaurant level orders are not compared to point of sale data, so it is unknown whether restaurant level adjustments of replenishment forecasts relate to actual end consumption. While deviation and bias will cause adverse effects in higher echelons of the supply chain regardless of this information, it would be useful to know whether the source is local information advantage or the result of some cognitive dissonance by the individual restaurant operator.

Conclusions

In this research, we examined the operational effects of governance form that, in part, addressed the claim by Combs et al. (2010) on the relative lack of research on

franchising operational performance after contract formation. As part of our inquiry, we incorporated the call of Waller and Fawcett (2013b) and Cotteleer and Wan (2016) to pair theoretical inquiry with big data exploration. This tactic helped protect against the risk of false positives inherent to big data exploration and of over-inflated statistical power when testing theory in big data. This also permitted rich contextualization of the effect of governance form on replenishment forecast deviation and bias. In doing so, we also developed a scalable method for examining and isolating effects in a hierarchical framework, called *HPD*. Our proposed method of analysis permitted rapid characterization of effects from millions of individual transactions with limited manual interpretation. This process, scalable to much larger populations, isolates temporal, regional, and product based peculiarities of the impact of governance form on our two dependent variables. Finally, our mixed results support previous findings on the effect of governance form on replenishment forecast deviation, but indicate greater theoretical work must be done to characterize and predict the effect of governance form on replenishment forecast bias. Future work must aim to identify the alternative causes or combinations of causes that drive higher bias in franchised restaurants.

Chapter 3: "The Impact of Including Forward Indicators on POS Demand Forecast Accuracy: The Case of Short-Term Weather Forecast Data"

Introduction

Demand forecasting attempts the unenviable task of predicting complex human preference and behavior with incomplete information. Short term demand predictions at the product level are most often executed through a statistical forecast developed from records of past demand, and extrapolated forward (Jain 2001, Fildes et al. 2015). Remarkably, this simple approach has managed to produce some highly accurate forecasts over wide ranges of industries, products, and locations, despite assuming stationarity of conditions which drive changes in demand (Armstrong 2002). Weather, an established driver of mood, preference, and ultimately consumer demand, has long been incorporated into statistical forecasts via estimates of decomposed seasonal effects. However, such estimates carry the implicit assumption that past seasons and their effects will occur again in the same way.

What seasonality estimates fail to capture is that weather can change drastically on a day to day basis, and has an immediate impact beyond a mean seasonal effect. This has motivated many forward-leaning firms to incorporate short-medium term weather forecasts into their demand planning processes. Increasingly accurate short-medium term weather forecasts are available to demand planners from a variety of sources, and

predictive weather indicators have increasingly shown promise in the last few years, though with some mixed results.

Nestle began incorporating weather forecast data into their bottled water demand forecasts in 2008, improving weekly sales forecasts by 2-6% (IGD 2009) and saving them as much as $12 million annually (Banker 2009). British grocery chain Tesco even started employing their own weather forecasters in 2009 to improve their demand forecasts (Werdigier 2009). Giants like Walmart and Proctor and Gamble paired with the Weather Company (owner of The Weather Channel and weather.com) in 2013 to match their point-of-sale (POS) data with weather data to identify trends down to the individual consumer level (Suddath 2014). IBM, recognizing a growing demand among businesses for accurate weather forecast data, purchased the sensor, digital and data assets of the Weather Company for $2 billion in late 2015 (Hardy 2015). Combining the Weather Company's assets with sensors from the National Oceanic and Atmospheric Administration (NOAA), NASA and the U.S. Geological Survey (Dillow 2011), IBM launched their Deep Thunder hyperlocal custom weather forecast engine in 2016, providing weather-based insights for their business clients (Stockton 2016). This expanding interest and investment in predictive weather indicators demands a greater academic investigation into implications of predictive weather indicators for business managers.

Observed weather's immediate effect on demand has been examined in a number of contexts and industries, with mixed conclusions on its nature, magnitude, and even direction within the academic literature (Bertrand et al. 2015, Arunaj and Ahrens 2016,

Bujisic et al. 2016, Tran 2016, Li et al. 2017, among the most recent). Though consensus exists that weather has an effect on demand, the causal linkages are still not comprehensively understood. Practically applying the limited understanding of the effects of weather on demand is further hampered by the fact that almost all research focuses on the effect of *observed* weather; information unavailable to demand planners when they make critical predictions for their supply chain. To date, despite growing use of predictive weather indicators among practitioners, little has been published on forecasting short term demand from *predicted* weather indicators (Nikolopoulos and Fildes 2013, Steinker et al. 2016).

This distinction between *observed* and *predicted* weather indicators is important, and is based in the inherent uncertainty of a weather forecast. Though the typical impulse for a forecaster is to include as much information as possible to improve accuracy, there is inherent risk in including information in a forecast which is itself uncertain. Thompson and Brier (1955), Thompson (1962), Murphy (1977), Katz and Murphy (1990) and Katz and Lazo (2011) all demonstrate, using cost-loss models, that imperfect weather forecast information can only improve expected economic value for a business decision maker if they are sufficiently reliable over the decision making horizon, and if it is possible to make investments that protect against negative weather effects. If a decision maker incorrectly estimates the reliability of the incorporated weather forecast, the cost or efficacy of a loss preventive investment, or the loss that would be associated with a weather event, they stand to lose money. Improper incorporation of uncertain information like weather forecasts places businesses in a wide array of industries at risk.

For instance in February of 2017, incorporation of an improperly specified temperature forecast that was only a few degrees off into a power load projection led to blackouts in over 40,000 South Australian homes in the middle of a dangerous heatwave (Burton 2017). Regarding economic risk, an estimated 16-25% of U.S. GDP and 80% of US companies are considered weather sensitive, or elastic to changes in weather (Bertrand and Sinclair-Desgagné 2011).

This risk generally increases as the uncertainty of the weather forecast increases, or as the horizon of the demand forecast is extended, but differs by application. Demand planners are interested in forecasting horizons, which at a minimum, extend through decision points where changes can be made to material flows (Murphy 1993). This is, of course, highly dependent on production and logistics lead time, as well as the degree of inventory and production centralization. For an industry like agriculture, there is no requirement to accurately predict weather on a day-to-day basis, but weather information is required months in advance. Climate forecasts indicate with reasonable accuracy the expected *accumulated* levels of rain, wind and sun a farmer might expect, and would dictate which fields they may fallow or which crops they may plant. For sales and operations planning, the required weather forecast information horizon is typically shorter, but requires much greater day-to-day accuracy. Weather effects can only be aggregated over the relevant planning horizon. For manufacturing concerns, weather forecasts would need to span a production cycle. To impact logistics costs, weather forecasts need to span a replenishment cycle.

Paradoxically, due to spatio-temporal aggregation mitigating the effects of short term variation, long-term climate forecasts tend to be more accurate than short and medium term weather forecasts (Camargo and Hubbard 1999, Janis et al. 2004). Short and medium term weather forecasts do not benefit from this aggregation effect, and so their accuracy significantly degrades at ranges past one-two weeks. Pepsi (France) experimented with incorporating weather forecasts into their sales and operations forecast in 2009, but ultimately abandoned the effort when they found weather forecast degradation past a two week horizon countered any benefit from inclusion (Fustier 2011). Their two-week production cycle was too long, or available weather predictions at the time were too unreliable to provide value to their forecasts.

In this study, we demonstrate the effect of including short-medium term predicted weather indicators in demand forecasts for the quick service restaurant industry. Utilizing autoregressive prediction models, we introduce exogenous weather variables into time series forecasts for 41 menu items at 2742 individual restaurants distributed throughout the continental U.S. We expand on initial work by Nikolopoulos and Fildes (2013), and Steinker et al. (2016) to estimate the effect of a greater variety of weather forecast variables, across more products and locations. In the process, we demonstrate actual improvement in various demand forecast quality measures through inclusion of predicted weather, and possible improvements as weather forecast reliability improves. Further, we demonstrate some instances where simple linear models that include predictive weather factors show improvement over the proprietary forecast generated by the restaurant firm over the same period. These improvements in forecast quality have

direct financial implications for the restaurant firm, and provide support for inclusion of readily available predictive indicators in forecasting efforts in broader contexts.

The remainder of the paper is divided as follows: first we review the literature regarding observed and predicted weather effects on demand and demand forecasts, next we describe our modeling effort, report comparative results from our models, discuss implications and draw conclusions from our results, present limitations, and finally highlight opportunities for future research.

Literature Review

Observed Weather's Effect on Demand

The economic effect of weather is well established in a number of familiar contexts, and there is a significant body of literature dedicated to describing it. The effects (and resultant implications) vary by industry, weather and demand forecast horizon, specific weather phenomena and scope or level of aggregation.

Though weather has sizable demonstrated effects on financial markets, manufacturing, retail, and services, the industries that have seen the most weather-related research are those that are most directly impacted by (and thus sensitive to) weather effects; agriculture and energy (Lazo et al. 2011). Agriculture depends directly on both immediate and accumulated climate effects. Agricultural papers typically model effects on crop yields (Mjelde et al. 1989, Potgeiter et al. 2003, Hamjah 2014), and depend on long-range climatological, rather than short-medium range weather forecasts for most predictive models. Energy related industries have an effect that is nearly as direct. Both mining and energy utility demand increase under conditions of higher energy

consumption. Consumption tends to increase with both high and low temperature extremes (Considine 2000, Auffhammer and Mansur 2014), and energy forecast models may incorporate either short term weather forecasts for load balancing or long-term climatological forecasts for capacity planning.

Restaurants are typically identified as part of the service industry (Howells and Morgan 2017), though at times are grouped with retail (Starr-McCluer 2000), and share multiple demand factors with retail. Lazo et al. (2011) notes a relative lack of research on weather sensitivity in the service industry sector, relative to agricultural and energy sectors. This is despite weather accounting for an estimated $60B in variation within service industrial sector revenue, compared to less than $16B in either agricultural or energy sectors (2008 dollars). Bujisic et al. (2016) cite a specific lack of research on weather sensitivity in hospitality and restaurant segments of the service industry sector outside of coarse climatic and seasonality effects. For this reason, we review the literature of weather sensitivity of both retail and services.

Weather Effects on Retail and Service Sectors

Steele (1951) was the first to demonstrate the effect of weather on retail sales, determining a negative impact from precipitation, snow accumulation, and ambient cooling on daily department store sales. Early studies of weather's effect on retail tended to be either limited in scope (like Steele's research to one store), or be regionally or temporally aggregated. This is likely due to data availability and computational limitations, and makes such studies of limited value for enterprise-level, short term and distributed demand planning. For example, Johnston and Harrison (1980) used monthly

nation-wide average temperature and sunshine deviation on aggregate UK cider sales. They found that increases in a combination of temperature and sunlight positively affected cider sales, but this analysis was not location specific. Juselius (1985) similarly used monthly nationwide averages, but of numbers of warm-temperature days, determining a significant positive effect on sales of Finnish soft drinks. Later studies in retail decreased the temporal or regional aggregation, or included additional effects.

Since these early retail studies, various specific weather effects, most notably temperature, have been empirically explored in relation to numerous contexts. Operationalizations of temperature have been found to have a nonlinear but generally positive effect on demand for a wide range of products, including lawn care products (Cawthorne 1998), aggregate nondurable products (Starr-McCluer 2000), soft drinks (Divakar et al. 2005, Ramanathan and Muyldermans 2010), beer (Bratina and Faganel 2008), aggregate service industry sales (Lazo et al. 2011), demand for both cars and homes with warm-weather features (Busse et al. 2012), online clothing (Steinker et al. 2016), and food and clothing retail sales (Arunaj and Ahrens 2016). This positive effect is, however, diminished or even negative in restaurant sales (Starr-McCluer 2000, Bujisic et al. 2016), winter sports demand (King et al. 2014), extreme temperatures (Parsons 2001, Tran 2016), or may be dominated by a negative effect from variation in temperature (Koksalan et al. 1999, Mena et al. 2014, Bertrand et al. 2015). The temperature effect can also be heterogeneous based on region (Divakar et al. 2005, Tran 2016), season (Johnston and Harrison 1980, Bahng and Kincade 2012), temporal position in a season (Cawthorne 1998, Choi et al. 2011), channel (Divakar et al. 2005), and

product (Starr-McCluer 2000, Choi et al. 2011, Busse et al. 2012, Arunaj and Ahrens 2016).

In addition to temperature, retail studies have found precipitation (rain or snow) to have negative impacts on demand in department stores (Steele 1951), outdoor malls (Parsons 2001), purchases via mobile phones (Li et al. 2017), online clothing (Steinker et al. 2016), food and clothing retail sales (Arunaj and Ahrens 2016), and sporting goods stores (Tran 2016). Though, this effect is reversed in demand for winter weather appropriate vehicles (Busse et al. 2012) or winter sports (King et al. 2014), and Lazo et al. (2011) note an overall positive effect in service industry revenues related to precipitation. Sunlight is found to reduce negative affect, increase demand for tea, coffee (Murray et al. 2010), alcoholic cider (Johnston and Harrison 1980), purchases via mobile phones (Li et al. 2017), and online clothing (Steinker et al. 2016), though not outdoor mall foot traffic (Parsons 2001), and exactly the opposite in demand for cars with winter weather features (Busse et al. 2012). Humidity reduces positive affect, and has a negative impact on restaurant sales (Bujisic et al. 2016), though is not found to significantly impact tea and coffee sales (Murray et al. 2010) or outdoor mall foot traffic (Parsons 2001). Wind also has been found to coincide with lower restaurant sales (Bujisic et al. 2016). As with temperature, these alternate weather effects tend to be both nonlinear and heterogeneous across a number of dimensions (Arunaj and Ahrens 2016, Tran 2016).

There are instances where we would not expect the effect of weather on quick service restaurant demand to resemble retail demand, such as with online (Steinker et al.

2016) or mobile (Li et al. 2017) purchases, or in purchases of some durable goods (Starr-McCluer 2000, Choi et al. 2011, Bahng and Kincade 2012, Busse et al. 2012, Bertrand et al. 2015). This is due to the experiential nature of restaurants. Though they sell tangible goods, those goods are typically consumed within a short time, and so are only purchased when there is an immediate need. This makes restaurant demand, particularly for the partially commoditized quick service restaurant industry, highly dependent on short term inclinations of people to be out and to physically visit a store. Steele (1951) posits four ways short term weather might affect a customer's desire (or ability) to visit a business.

Explanations for Consumer Behavior

First, they may be physically prevented, as would be the case in an extreme weather event. Second, a shopper may be disinclined due to inconvenience. This might be the case with severe cold, heat, fog, snow or precipitation, which would require additional planning, protective clothing, or caution. This notion is supported by research that links increased outdoor leisure activity to increases in temperature (Smith 1993), though the effects are regionally and seasonally heterogeneous (Tucker and Gilliland 2007), and negative in extreme temperatures (Zivin and Neidell 2014). These first two mechanisms are facilitated by local infrastructure and individual adaptability, driven by local weather norms (Tran 2016). Third, the shopper may face psychological barriers to either go out, or once out, to make a particular purchase. There is a tremendous amount of psychology and marketing literature which indicates linkages of weather with mood, and mood with behavior (Cao and Wei 2005). Persinger and Levesque (1983) indicate that 40% of mood evaluations can be attributed to weather. Increased temperature

(Cunningham 1979, Howarth and Hoffman 1984), sunlight (Cunningham 1979), barometric pressure (Goldstein 1972), and decreased humidity (Sanders and Brizzolara 1982, Murray et al. 2010) all relate to positive affect; with negative affect mitigated by sunlight (Murray et al. 2010). Positive affect is then positively related to purchase interaction quality (Gardner 1985), positive perceptions of goods (Bitner 1992), and increased spending (Donovan et al. 1994, Spies et al. 1997); with negative affect related to a decreased willingness to pay (Murray et al. 2010). Some research also links "bad" weather (increased precipitation, fog, and decreased sunlight) to risk-averse behavior (Hirshleifer and Shumway 2003, Bassi et al. 2013, Li et al. 2017). Fourth, a shopper may perceive different product utilities based on weather conditions. Unexpected rain may spur the purchase of ponchos or umbrellas, and an early snow flurry may initiate the season for selling winter garments. Subsequent studies (Starr-McCluer 2000, Tran 2016) have measured this by incorporating weather variables into household production utility models. This may also manifest as weather effect heterogeneity, and in cases of substitution where weather has individual product effects, but not overall demand effects (Choi et al. 2011, Bahng and Kincade 2012).

Problems with Using Observed Weather as a Proxy for Weather Forecasts

Previous work relating past observed weather effects to demand were either descriptive, in that they did not claim to be able to predict future behavior with the identified relationships, or they were implicitly forecasting weather along with demand. Murphy (1997) and Armstrong (2002) refer to this as ex-post or conditional forecasting, and note that while it can provide extremely accurate description of past behavior, it can

perform quite poorly when predicting behavior. The problem with this approach is that the best known methods for statistically estimating future demand from a time series differ substantially from the best known methods for forecasting weather.

Advances in Weather Forecasting

It is useful at this point to define what is meant by a short term or medium term forecast. The National Weather Service (NWS) and American Meteorological Society define short term forecasts as up to two days ahead, and medium range forecasts as being between two and seven days ahead (AMS 2015). We use this definition for short range weather predictions, though as has increasingly been the case in recent years (Hu and Skaggs 2009) we extend the definition of medium range weather prediction out to ten days.

Short and medium range weather forecast accuracy has increased rapidly in recent decades, and the reason for this is also the primary reason why it is advantageous to estimate weather separately from demand. While demand forecasting is primarily limited (at least mathematically) to statistical and probabilistic extrapolations, weather has (for quite some time) been better estimated through an ensemble of methods that include simulation. From the earliest manual attempts by Lewis Fry Richardson in 1922, to the more successful computerized efforts in the 1940s by the mathematician John von Neumann, large scale simulation of fluid mechanic and thermodynamic weather effects have accuracy limited only by computing power and environmental sensor data availability (Tribbia 1997). Significant public and private investments in environmental sensors and exponential advances in computation have permitted steady improvements in

weather forecast accuracy via Monte-Carlo simulation based ensemble forecasts (Dutton 2002). NWS short term temperature forecast average error was cut in half between 1966 and 2014, now between 2.5-3.0^{o}F for two-day ahead forecasts (Huntemann et al. 2014). Between 1992 and 2012, temperature forecasts of five-six days achieved the previous accuracy of three-four day forecasts (AMS 2015). Probabilistic forecast performance for short and medium range precipitation have improved similarly (Hu and Skaggs 2009, Huntemann et al. 2014). Overall, the reliable forecast range of most weather phenomena has increased roughly one day each decade, and is currently greater than one week (AMS 2015).

Effort has also been made to forecast demand using simulation, but demand planners still primarily rely on extrapolative time series methods for quantitative forecasting (Fildes et al. 2015). The reason for this is that the required econometric and behavioral simulation parameters for demand forecasting depend on much less reliable information than Newtonian factors which are found to drive short term weather effects. Forecasters have found limited success extrapolating exogenous effects in what Armstrong (2002) terms static simulation, which supposes that an effect will not change from period to period. Similarly, judgmental adjustment bootstrap simulations outperform both manual adjustments and forecasts without subjective adjustments, but only under a narrow range of stable exogenous conditions (Ritzman and Sanders 2001, Fildes et al. 2008). This is clearly not a suitable assumption for most daily weather conditions, and fails to leverage the tremendous advances in predictive power stemming from an accurate understanding of physical forces, simulation, exponentially increasing

computing power, and an ever expanding network of physical sensors. Therefore, despite so many previous efforts to extrapolate weather and demand concurrently, we choose to incorporate separately generated weather forecasts as exogenous factors in a statistical model for demand.

Forecasted Weather to Predict Demand

Although observed weather's effect on demand is widely studied, if somewhat less-so in restaurant and service contexts, it is of limited utility for prediction. As all of the above listed research measures the effect of only observed weather, they imply an ability for perfect weather prediction. We must acknowledge that information available to demand forecasters is imperfect, and so should include accurately predicted weather information rather than perfect observed weather information to estimate predictive models. This is particularly true in a supply chain context, where immediate decisions about sourcing, production, and material flow relate to sales and operations plan with horizons of a week or more. The director of the NWS recently noted that weather forecast accuracy drops off considerably after a few days, due to unpredictable lower order effects from physical inputs (Palmer 2013). While the NWS and many other providers now offer 10-day forecasts, and there is evidence of forecaster skill that extends as far as 14 days (Stern and Davidson 2015), accuracy beyond that range tends to be no better than what you may find in the Old Farmer's Almanac (Samenow and Fritz 2015). Silver (2012) notes that most temperature forecasts beyond nine days in advance actually tend to perform *worse* than historical averages, and only slightly better than a naïve persistence forecast. This limitation has in the past made weather forecast information

less valuable for businesses with longer production cycles or lead times, and certainly calls into question the use of observed weather as a proxy for predicted weather when demonstrating the effect of incorporating weather into demand forecasts.

The work of Thompson and Brier (1955), Thompson (1962), Murphy (1977), Katz and Murphy (1990) and Katz and Lazo (2011) indicate that weather forecast reliability over a business's operational planning horizon is a necessary condition for it to provide value. Weather forecasts with sufficient lead time to permit changes in material flow or capital investments have only in recent years become more reliable than simple historical trends or seasonality estimations (Stern and Davidson 2015), a requirement for use in decision-making (Mjelde and Dixon 1993). To date, only two studies have observed the impact of incorporating medium range weather forecasts on demand forecast accuracy.

Nikolopoulos and Fildes (2013) published the first work that evaluates the effect of incorporating medium range weather forecasts in a demand forecast. Highly context-specific, they investigated the effect of including 10-day ahead temperature deviation predictions in demand forecasts of beer sales in the United Kingdom. They indicate significant improvement in demand forecast accuracy by this inclusion, especially in warmer months of the year. These results also support previous indications that weather effects vary by region, season, product, mood, and a number of other difficult to capture factors. Steinker et al. (2016) expand on this initial inclusion of weather forecasts, but in the (again) specific context of German online retail in only two cities. They examine the effect of including seven-day ahead predicted sunlight, temperature and precipitation on

demand forecast accuracy. They first establish an upper-bound fitting a model to in-sample data with observed weather, then validate the model in an out-of-sample forecast generated with historical weather forecast data. They find sunlight and temperature are positively related to online sales, with this effect increased on the weekend, whereas rain is negatively related. This is consistent with expectations for weather impacting the time spent indoors versus outdoors. They also note a significant improvement in demand forecast accuracy through the inclusion of weather forecast indicators, though the effect diminishes as forecast horizon increases.

In sum, previous research on the effect of observed weather on demand have indicated significant effects from a number of weather variables in a variety of industrial and regional contexts, though this effect tends to be heterogeneous across industries, regions, seasons, temporal positions in a season, and product. Research also indicates that predicted weather can improve demand forecasts, but to date this is limited to a narrow scope and context. We wish to expand on the scope of previous work and test the potential of weather forecast data to improve demand forecast quality over a broader set of contexts and in a new industrial setting.

Data

Our data were collected from multiple sources. Historical time series demand data was furnished by the primary fourth party logistics (4PL) provider for a major international quick service restaurant. Through an ongoing relationship with a university research group, they enlisted our assistance in helping determine the effect of predictive indicators which may improve their forecasting. To accomplish this, they provided us

with a large sample of forecasts and corresponding POS records they considered to be representative of a wide range of menu offerings, demand patterns, geographic regions, and seasons. The potential sample included 41 of the most popular menu-level products at 4240 individual restaurants, covering a date range from 26 September 2014 through 1 February 2016. In total, we evaluated nearly 85 million individual observations, with each corresponding to an individual menu item-location-day.

Demand planners at the 4PL currently generate rolling POS and replenishment forecasts once each week for each menu item-location. If weather forecast information is to be included in these POS forecasts, it must have a horizon of at least seven days, and cover the same geographic regions and time periods as our POS sample. Though the NOAA does not systematically store NWS weather forecasts, we were able to obtain The Weather Company's historical weather forecast data from a third party weather forecast monitoring and assessment firm called ForecastWatch. The firm gathers forecasts from multiple public and private sources, and regularly assesses their relative performance. Most private weather forecasters use the NOAA network and even the NWS forecast as a basis for improvement (Silver 2012), but they vary in how much (if at all) they improve on the public forecasts. In a 2014 assessment, The Weather Company's one-nine day temperature forecasts were competitive with the best domestic forecast provider (WeatherUnderground) and better than the NWS forecast (Floehr 2015).

We collected daily data from 836 available airport weather stations in the continental US, providing widespread geographic coverage and tracked by International Civil Aviation Organization (ICAO) codes. Each daily prediction includes a nine, five,

three and one day prior forecast for high and low temperature, vector average wind speed, five point Likert scale for cloud cover percentage, and five point Likert scale probability of rain, thunderstorms, snow, and overall precipitation. ForecastWatch provided daily observed high and low temperature, vector average wind speed, and accumulated precipitation for each observation. We augmented their data with observed daily rain, thunderstorm and snow occurrence for each ICAO code from the NOAA's Local Climatological Database (NCEI 2017). We include multiple measures of predicted weather indicator reliability for all point forecasts (temperature and wind speed) in Table 1. Mean error (ME) is an indicator of bias and mean absolute error (MAE) is a measure of accuracy for point forecasts. Bias decreases with increased forecast horizon, as the further a projection is, the more closely it resembles long-term climatology. There is also an indication of conservatism in the predictions, as bias (for all three phenomena) is directionally away from extreme values. Accuracy degrades with longer horizons, which may negatively affect the value of predicted indicators in longer horizons.

Point Forecasts	*One-Day*	*Three-Day*	*Five-Day*	*Nine-Day*
		ME		
High Temp (^oF)	-0.64	-0.64	-0.56	-0.41
Low Temp (^oF)	0.32	0.35	0.24	0.07
Wind Speed (mph)	-2.46	-2.20	-1.97	-1.58
		MAE		
High Temp (^oF)	2.22	2.89	3.78	5.75
Low Temp (^oF)	2.28	2.84	3.61	5.05
Wind Speed (mph)	2.90	2.88	3.10	3.62

Table 1: Predicted Weather Indicator Point Forecast Reliability

Table 2 includes measures of predicted weather indicator quality for all probabilistic forecasts (rain, thunderstorms, snow and overall precipitation). The Brier Score is often used to assess probability forecasts, and is a special case of mean squared error (MSE) bound by zero and one (Murphy 1997). As with all cases of MSE, lower values indicate better forecasts. Unfortunately, this measure equally rewards correctly predicting both occurrences and non-occurrences of a weather event. For rarer events like snow or thunderstorms (that may have a significant effect on demand), this value is artificially low. We, therefore include a measures of positive predictive value (PPV) and sensitivity (Brenner and Gefeller 1997). PPV expresses the proportion of positive occurrences given positive predictions of an event. For probabilistic forecasts, we use the classification probability cutoff of 0.5, indicating that an event is predicted to be more likely than not to occur. Sensitivity expresses the proportion of positive predictions given positive occurrences of an event. As with the point forecast quality indicators, the probability forecast quality indicators generally degrade with longer horizons. PPV and

sensitivity also indicate conservatism in predictions, as the predicted probabilities and frequencies of weather events tend to be lower than the observed probability. For example, 94% of one day ahead rain forecasts predicting a 50% or greater chance of rain are followed by an observed occurrence of rain, but only 23% of rain events are predicted.

Probability Forecasts	***One-Day***	***Three-Day***	***Five-Day***	***Nine-Day***
		Brier Score		
Rain	0.306	0.309	0.326	0.365
Thunderstorm	0.067	0.072	0.073	0.072
Snow	0.035	0.041	0.044	0.054
Precipitation (all)	0.176	0.179	0.199	0.238
		PPV		
Rain	0.936	0.878	0.804	0.586
Thunderstorm	0.375	0.336	0.298	0.250
Snow	0.887	0.720	0.648	0.377
Precipitation (all)	0.956	0.889	0.816	0.595
		Sensitivity		
Rain	0.228	0.239	0.215	0.208
Thunderstorm	0.512	0.563	0.491	0.282
Snow	0.333	0.301	0.265	0.127
Precipitation (all)	0.405	0.421	0.359	0.248

Table 2: Predicted Weather Indicator Probability Forecast Reliability

One limitation of our available weather predictions are the gaps in projection (i.e. one, three, five and nine rather than one through nine day predictions). As a result, we "bin" effects from projected weather into horizon categories. In each weekly demand forecast, six-seven day horizons depend on weather projected nine days prior, four-five

on five days prior, two-three on three days prior, and next day symmetrically matched. The resultant conservative application of proxies for full weather forecasts means our results will serve as a lower bound estimate for weather accuracy at longer ranges.

Relevance of station forecasts to individual restaurants was also an area of concern. In order to sufficiently capture geographic weather variance in the U.S., minimum sensor density is dependent on the weather measure of interest. Camargo and Hubbard (1999) found that distances could not exceed 60 km in order to capture 90% of inter-site variation in daily max temperature. This distance reduces to 30 km for min temperature and sunlight, 10 km for wind, and five km for precipitation. Micro-climates in mountainous areas can drive the minimum distance as low as one km. Unfortunately, many of the 836 weather stations we gathered data from were too far distant from the nearest restaurants in our sample to be relevant by this metric. In this case, we compromise granularity of the measure with relevant sample size. We use the conservative threshold of 30 km distance to identify a weather station as being relevant to a restaurant. This ensures that temperature effects can be accurately estimated, and that a less granular categorical characterization of wind and precipitation can be used. Applied as the geodesic ellipsoid distance threshold between coordinates for stations and restaurants (NGIA 2014, Hijmans et al. 2016), this eliminated about 25% of our original restaurant sample (3162 remaining from 4240), and disproportionately from more sparsely populated Western states.

Hypotheses

As indicated in Nikolopoulos and Fildes (2013) and Steinker et al. (2016), we expect predicted temperature variables to have a significant (if heterogeneous) effect on demand, and so inclusion is likely to improve demand forecast accuracy. The direction of the weather effect does not matter when estimating forecast models, so the heterogeneity observed across several dimension (Johnston and Harrison 1980, Cawthorne 1998, Starr-McCluer 2000, Divakar et al. 2005, Choi et al. 2011, Bahng and Kincade 2012, Busse et al. 2012, Arunaj and Ahrens 2016, Tran 2016) will not affect the accuracy of a prediction using weather predictors, providing the effect is correctly estimated. By including both daily high and low temperature data, we capture the expected effects from the most extreme possible values. During summer months when high temperatures are more likely to have an effect, the high temperature indicator is more likely to have an effect on demand. In winter months, the low temperature indicator is more likely to have an effect on demand. Temperature is also one of the more reliably estimated weather parameters over a short horizon (Camargo and Hubbard 1999), but accuracy degrades at middle ranges (Floehr 2015). This leads to the following hypotheses:

H1a: Demand forecast models that incorporate exogenous high temperature predictions will be more accurate than models that do not.

H1b: Demand forecast models that incorporate exogenous low temperature predictions will be more accurate than models that do not.

H1c: Demand forecast models that incorporate short range (one to three days) exogenous temperature predictions will be more accurate than models that incorporate medium range (five to nine days) exogenous temperature predictions.

Bujisic et al. (2016) indicate wind has a significant negative effect on restaurant demand, and so inclusion is likely to improve demand forecast accuracy. Wind prediction, though also quite accurate (Camargo and Hubbard 1999), degrades in quality with longer forecast horizons. We therefore hypothesize the following:

H2a: Demand forecast models that incorporate exogenous wind speed predictions will be more accurate than models that do not.

H2b: Demand forecast models that incorporate short range (one to three days) exogenous wind speed predictions will be more accurate than models that incorporate medium range (five to nine days) exogenous wind speed predictions.

The effect of sunlight on negative (Murray et al. 2010) and positive (Cunningham 1979) affect is well established, and it has been found to positively relate to willingness to pay (Murray et al. 2010), interaction quality (Gardner 1985), perceptions of goods (Bitner 1992), decreased risk aversion (Li et al. 2017), and increased spending (Donovan et al. 1994, Spies et al. 1997). However, it is still unclear whether this translates to customer propensity to physically patronize businesses (Parsons 2001). Based on previous work that indicate a significant effect of this weather phenomenon on demand, we predict its inclusion will improve demand forecast accuracy. We do not have an indicator of weather forecast quality for sunlight (cloud cover), so will not speculate differences in forecast horizon. The resulting hypothesis is:

H3: Demand forecast models that incorporate exogenous cloud cover predictions will be more accurate than models that do not.

Various forms of rain and other precipitation have been found to have a significant, but heterogeneous, effect on demand (Parsons 2001, Lazo et al. 2011, Busse et al. 2012, King et al. 2014, Arunaj and Ahrens 2016, Steinker et al. 2016, Tran 2016, Li

et al. 2017). As posited in Steele (1951), this is likely negative (and more pronounced) when more extreme weather such as snow and thunderstorms make business patronage inconvenient or impossible. Unfortunately, precipitation forecasting tends to have lower reliability, particularly for locations further from a weather station (Camargo and Hubbard 1999). This is especially true as the forecast horizon increases. As a result, we hypothesize:

H4a: Demand forecast models that incorporate exogenous rain predictions will be more accurate than models that do not.

H4b: Demand forecast models that incorporate exogenous thunderstorm predictions will be more accurate than models that do not.

H4c: Demand forecast models that incorporate exogenous snow predictions will be more accurate than models that do not.

H4d: Demand forecast models that incorporate exogenous precipitation (all kinds) predictions will be more accurate than models that do not.

H4e: Demand forecast models that incorporate short range (one to three days) exogenous rain, thunderstorm, snow, or precipitation (all kinds) predictions will be more accurate than models that incorporate medium range (five to nine days) predictions.

Methodology

Autoregressive Models

To as great an extent as practical, we replicated the 4PL firm's forecast conditions. This was to ensure enhanced comparability with their own forecasting efforts. We did not have access to the forecast management system in use by the firm, and so could not replicate individual forecast model decisions. The firm customizes models, parameters, and adjustments in some cases to even the product-restaurant level using a combination of quantitative extrapolation and subject matter expertise of local

conditions. By comparison, we use a single (albeit adaptive) method to generate all forecasts, but include weather forecast indicators that are not included in the 4PL's models. Despite our inability to directly compare identical models, it is interesting to observe whether the inclusion of predictive indicators can improve a generic time series model to the extent that it might even outperform a more customized model.

We wished to generate POS forecasts for each product, at each restaurant, for each day over a seven day horizon, once each week. This required that we generate as many as 174,000 separate sets of rolling forecasts. Each set of rolling forecasts would be replicated with each of eight weather effects under both perfect knowledge and uncertainty, and include on average 10-15 re-estimated rolling forecasts. It was apparent that an automated method of forecast generation was required. Automated forecasting methods have been shown to outperform methods with static manual estimation, and adaptive to multiple time series characteristics (Makridakis and Hibon 2000). In addition, we needed a method that permitted effective control of exogenous effects. We wished to observe the effect of including exogenous short term predicted weather variables. In order to isolate these effects in a regression model, we had to include corrections for violations of ordinary least squares regression. Specifically, we wish to correct for autocorrelation, trend and seasonality if they exist. ARIMA is a flexible class of model that can incorporate these corrections and include exogenous regressors, all in an automated fashion.

ARIMA models include autoregressive (AR) terms, or lagged values included as predictors of outcome variables. They include differencing or integration (I) terms, to

transform non-stationary time series to be stationary. They also include moving average (MA) terms that express errors as linear combinations of current and lagged errors. Each of these terms have positive integer orders denoted by *p*, *d*, and *q*, indicating the number of lagged, differencing, and moving average terms respectively that the outcome variable is regressed on. To account for seasonality, ARIMA models can be adapted to include additional lagged, moving average, and differencing terms, denoted by *P*, *Q*, and *D* respectively with backshift operators in multiples of *m* seasons. This model, denoted ARIMA$(p, d, q)(P, D, Q)_m$, is expressed as (Cools et al. 2009, Arunaj et al. 2016):

$$\phi_p(B)\Phi_P(B^m)(1-B^m)^D(1-B)^d Y_t = \theta_q(B)\Theta_Q(B^m)\varepsilon_t$$

Where Y_t is a time series observed value (daily sales) at time *t*, $\phi_p(B)$ and $\Phi_P(B^m)$ are the non-seasonal and seasonal autoregressive operators with respective orders *p* and *P*, $\theta_q(B)$ and $\Theta_Q(B^m)$ are the non-seasonal and seasonal moving average operators with respective orders *q* and *Q*, $(1-B)^d$ and $(1-B^m)^D$ are the non-seasonal and seasonal differences with respective orders *d* and *D*, and ε_t is a residual error term at time *t*. All operators of form $\alpha_x(B)$ represent polynomials of backshift operators of form: $1-\alpha_1(B)-\alpha_2(B^2)-\ldots\ \alpha_x(B^x)$.

To include the effects of exogenous variables, the seasonal ARIMA terms can be represented as the stochastic error term in a multiple linear regression model (Aburto and Weber 2007, Cools et al. 2009, Peter and Silvia 2012, Arunaj et al. 2016):

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{i,t} + \eta_t$$

Where Y_t, the dependent variable, is a time series observed value (daily sales) at time *t*, *X* represents *k* separate independent regressors, β represent multiple regression coefficients of the *k* regressors, and η_t is a residual series defined by the seasonal ARIMA parameters:

$$\eta_t = \frac{\theta_q(B)\Theta_Q(B^m)}{\phi_p(B)\Phi_P(B^m)(1-B^m)^D(1-B)^d}\varepsilon_t$$

This model, referred to as seasonal ARIMA with external regressors (SARIMAX), was estimated for each menu item-location combination. Since each time series would likely exhibit distinct seasonality, trend, and autoregressive characteristics, we chose an adaptive algorithm developed by Hyndman and Khandakar (2008), the 'auto.arima' function in the 'forecast' package v.8.0 in R (Hyndman et al. 2017).

As described above, when external regressors are used, ARIMA parameters are estimated from the residuals of a linear model predicting the time series of interest. The 'auto.arima' function estimates seasonal ARIMA parameters automatically by, first, selecting the order of differencing using successive unit root tests (Hyndman and Athanasopoulos 2014). As Hyndman and Khandakar (2008) describe, it estimates seasonal difference order (*D*) using the Osborn, Chui, Smith and Birchenhall unit root test. This has been shown to perform favorably compared to other unit root tests for seasonal differences (Rodrigues and Osborn 1999). Second, it estimates non-seasonal difference order using the Kwiatkowski-Phillips-Schmidt-Shin test, as it corrects a bias toward over-differencing found in other tests of stationarity by testing the assumption that $d = 0$, rather than $d = 1$. Once *D* and *d* are determined, it estimates values of *p*, *q*, *P* and

Q by minimizing the corrected Akaike's Information Criterion (AICc), a correction of the AIC for finite samples:

$$AICc = -2\log(L) + 2(r) + \frac{2(r)(r+1)}{(n-r-1)}$$

Where *L* is the maximum likelihood function value for the model, *n* is the number of observations used to fit the model, and $r = p + q + P + Q + k + 1$ is the number of model parameters. Model parameters include all order parameters, the variance of the random error term ε_t, and $k = 1$ when there exists bias in the ARIMA error, else $k = 0$. AICc rewards goodness of fit and penalizes both model complexity and small relative sample size, resulting in high performing, but parsimonious models (Hyndman and Athanasopoulos 2014).

Models for Comparison

In order to measure the effect of including predicted weather variables in seasonal ARIMA models, what we will call the *evaluation* models, we include three benchmarks as a mean for comparison. First are the *baseline* models, with no external predicted weather indicators. This is the most direct method of comparison. Second, following the example of Steinker et al. (2016), we generate *upper-bound* models that include observed (rather than predicted) weather external covariates to forecast demand over the evaluation period. This is obviously unavailable to the demand forecaster, but provides an idea of the potential for weather's inclusion as weather prediction continues to improve. The third comparison, and least direct, are the *4PL* forecasts. The *baseline*, *upper-bound* and *4PL* forecasts are either generated or collected for each menu item-location for comparison against the *evaluation* models.

We begin by estimating *baseline* seasonal ARIMA forecast models without exogenous variables for each menu item-location combination. Each *baseline* model is fit on a full year of in-sample POS data to capture all potential annual seasons. This exceeds the requirements suggested by Hyndman (2007) of $p + q + P + Q + d + mD + 1$ minimum observations required to estimate a seasonal ARIMA model, which is based on statistical estimability only. This does not take into account the potential for increased model fidelity achievable when all annual seasonal variations are included in a training set. This stringent requirement reduced our sample further, as not all menu item-locations had the minimum training sample size.

The estimated model is then used to forecast over a seven day horizon. The out-of-sample data, or all observations that occur after the minimum 365 days of in-sample data, is then used for generating a successively updated forecast. This is a common method of forecast validation (Armstrong 2002), but we augment this further by also re-estimating model parameters with updated POS data. In seven day intervals, all model parameters are re-estimated with the previous 365 days of POS observations. In this way, we generate forecasts over the entire test period with the same frequency that the 4PL firm would, and ensure models have the best possible fit over a forecast horizon. Of the retained menu item-location combinations, re-estimation was conducted an average of 10-15 times.

Models Including Exogenous Observed and Predicted Weather Indicators

We include weather effects individually for four reasons. First, although we can assess the relative significance of an external regressor in a seasonal ARIMA model, we

cannot know whether a significant predictor improves demand forecast quality. Therefore, we must assess each regressor separately to determine whether it improves predictive power. Second, each weather effect varies in its reliability. It is useful to separately observe differences in demand forecast improvement from inclusion of factors that differ in reliability. Third, this helps to prevent manual heterogeneous variable selection. As we saw in previous efforts to quantify the effect of weather factors on demand, the effects were heterogeneous over a number of dimensions. Each menu item-location combination may indicate significance of separate weather factors that requires either manual model fitting or potentially misspecified models. For thousands of separate forecasts, this would not be practical, and would limit comparability of each forecast. Fourth, and perhaps most limiting when incorporating external regressors in an autoregressive model, high covariance among exogenous regressors makes a model inestimable. This is of particular concern with weather effects that tend to be highly correlated. Wind and rain, for instance, almost always coincide with thunderstorms, and snow and low temperatures are strictly linked.

Just as in the *baseline* models, each model with exogenous observed and predicted weather is fit on a full year of in-sample POS data. In-sample observed weather phenomena is included to capture the effect of weather. As mentioned previously, observed cloud cover (sunlight) data was unavailable, so we include one day ahead forecasts as a proxy. One day ahead forecasts are presumed to be the most accurate available substitute for observed data.

Each model is then projected forward with out-of-sample POS data. For the *upper-bound* forecasts, this includes out-of-sample observed weather variables as covariates. *Evaluation* models instead include out-of-sample predicted weather over a seven day horizon. As with the *baseline* model, this forecast occurs every seven days, and each successive forecast re-estimates model parameters with the previous 365 days of in-sample POS and observed weather data.

Forecast Evaluation

From each successively updated forecast, we generate indicators of forecast quality. As Hyndman and Koehler (2006) note, each proposed measure of forecast quality has limitations. Some are scale-dependent, such as MAE or root mean squared error (RMSE). These are more useful for calculating costs, but provide no value for comparison of forecasts. Measures based on percent error, such as mean absolute percent error (MAPE) are popular in practice because they are easy to calculate and can be useful for comparison. This is also the measure currently in use by the 4PL. Unfortunately, MAPE suffers from inflation in time series with low values and is inestimable for time series with zero values. Because of these deficiencies, and because both Makridakis and Hibon (2000) and Armstrong (2002) recommend using multiple measures of accuracy, we also include the bias indicator mean error (ME), the relative error indicator Theil's U (Theil 1966), and mean absolute scaled error (MASE), a scaled error term suggested by Hyndman and Koehler (2006). MASE removes scale by comparing a forecast to a known method (usually naïve), and also permits multiple period out-of-sample error estimation, not possible with relative error measures like Theil's U. Error is scaled using

in-sample naïve forecast MAE (Hyndman and Koehler 2006) $q_t = \frac{e_t(h-1)}{\sum_{t=2}^{h}|Y_t - Y_{t-1}|}$ for nonseasonal data, or in-sample seasonal naive forecast MAE (Hyndman and Athanasopoulos 2014) $q_t = \frac{e_t(h-m-1)}{\sum_{t=m+1}^{n}|Y_t - Y_{t-1}|}$ for seasonal data, where $e_t = Y_t - F_t$ is the difference between forecast and observed demand, *m* is seasonality and *h* is the length of the forecast horizon. MASE is then simply $\sum_{t=1}^{h}|q_t|/h$. Of note, though we generate daily forecasts, MAPE is generated over the relevant replenishment period to limit inflation and ensure estimability. Restaurant replenishment occurs twice weekly, so each seven-day forecast period is evaluated on two replenishment periods. We finally also include weekly MAPE as this is how the 4PL firm tracks accuracy, so that a direct comparison can be made.

Addressing Computational Scale

Automated parameter estimation permits an increase in the number of menu item-location combinations to be forecast by limiting manual model estimation. However, even after filtering out menu item-locations with too few observations or at distances too far for reliable weather effect estimation, we estimated 105,875 *baseline* sets of forecast models. After matching available observed and predicted weather data, we estimated an additional 747,542 *upper-bound* and *evaluation* sets of forecast models. Finally, we calculated forecast quality metrics for each individual seven day forecast model contained in each rolling set, as well as for each individual seven day forecast period provided in the *4PL* sample. In total, we generated and (or) evaluated over 1.7 million sets of rolling forecasts. Particularly since we re-estimated ARIMA parameters in

successively updated forecasts, this became computationally burdensome. As a result, utilized extensive parallel computing for tractability. All forecasts were generated through the Ohio Supercomputer Center's Owens Cluster, a 23,392-core HP Intel Xeon E5-2680 v4 machine capable of 750 teraflops (OSC 1987).

Output Analysis

In order to determine the effect of including predicted weather variables on demand forecasts, we compare the eight generated forecast quality measures in separate linear models. We include a categorical predictor of external weather variable used to generate a forecast. Each of the aforementioned eight weather prediction variables represents a factor level, with the *baseline* condition serving as the reference level.

We control for two expected continuous sources of variance in each of the models. Increased volatility in times series is known to inherently reduce forecastability, so we control for this by including the coefficient of variation (CV) as a covariate (Armstrong 2002). Our method of matching restaurant locations via a single distance threshold also likely has an effect on weather forecast accuracy, and thus demand forecast quality derived from weather. Both observed and predicted weather relevant to each ICAO weather sensor will more accurately reflect the weather conditions at restaurants close by, but less so at restaurants close to the cutoff threshold. We therefore control for potential differences by including geodesic ellipsoid distance as a covariate.

In addition to continuous predictors likely to influence accuracy, we also control for two expected categorical sources of variance. To account for confounding effects from regional and product-based heterogeneities in weather variable effects on measures

of demand forecast quality, we separate our analysis by product categories and regions. As we have no hypotheses regarding the direction or size of the heterogeneous effects, we chose a weighted effects coding scheme that merely measures differences from an overall mean (Darlington and Hayes 1990, Cohen et al. 2013).

Regional heterogeneities are accounted for using weighted effects codes for the nine climatic regions defined by the NOAA as exhibiting similar characteristics for temperature and precipitation for more than a century (Karl and Koss 1984).

Figure 3 (Sanchez-Lugo 2017) displays the regional boundaries used in this research, and relative frequency of restaurants for each climate region are reported in

Table 3. While more granular microclimate divisions exist that could account for greater degrees of weather differentiation (Vose et al. 2014), these more aggregate regions have more explainable differences in response to specific weather effects. For instance, it is likely that consumer response to snow in the Northeast, where such weather is common winter months, will not be the same as in the warm Southwest. However responses in Albany and Buffalo, NY, who occupy different microclimate divisions (Fenimore 2017) but the same climate region, would likely not differ substantially due to similar weather patterns.

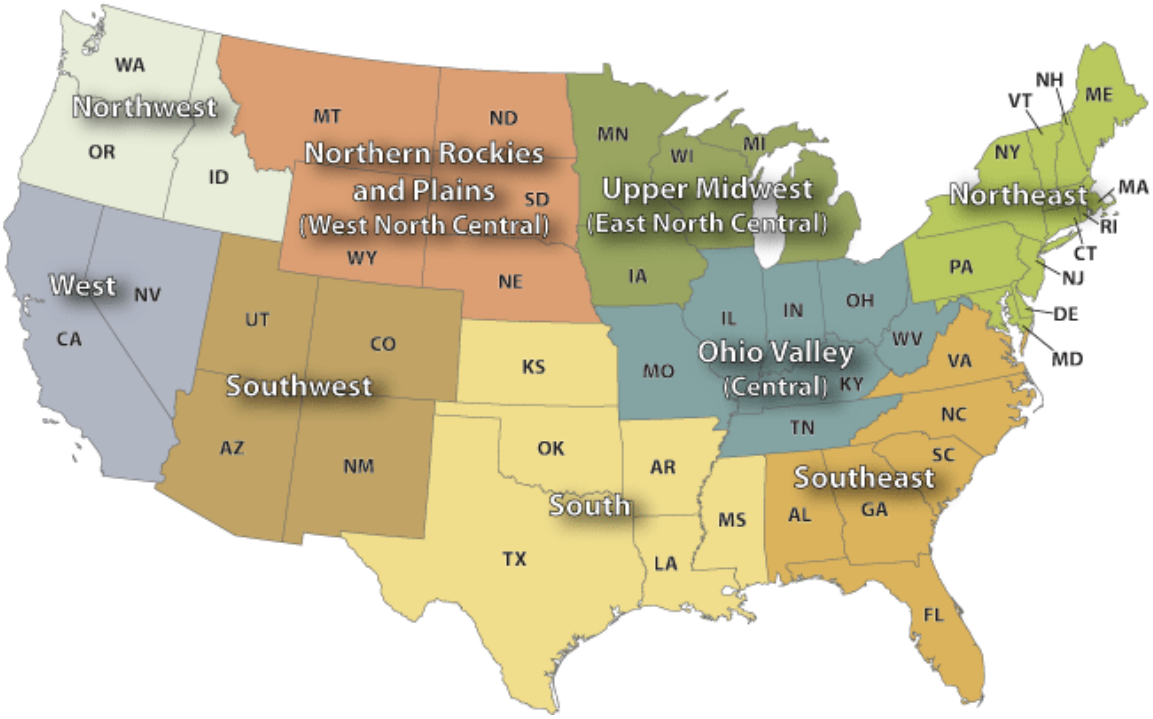

Figure 3: NOAA Climate Regions

NOAA Region	*No. of Restaurants*
Central (Ohio Valley)	663
East North Central (Upper Midwest)	4
Northeast	300
Northwest	0
South	583
Southeast	375
Southwest	6
West	758
West North Central (Northern Rockies and Plains)	53

Table 3: Restaurant Sample by NOAA Region

We account for potential product-based demand response heterogeneities similarly to Bujisic et al. (2016). They separate their full service menu items as main courses, sides, as children's or adult meals and by mealtime. We coded menu items as breakfast entrees or sides, lunch/dinner entrees or sides, desserts or shakes, drinks, and add-ons or specialty items.

Results

Comparing output from the four sets of models (*evaluation, baseline, upper-bound* and *4PL*), we find some interesting results. We began with a preliminary comparison of means for each of the forecast quality measures we generated as shown in Table 4.

	ME	*RMSE*	*MAE*	*MASE*	*Theil's U*	*MAPE 1*	*MAPE 2*	*Weekly MAPE*
4PL	-1.132	14.903	12.269	*	0.838	25.538	27.330	20.688
Baseline	-1.514	18.095	15.015	0.860	0.908	18.302	19.473	14.932
Upper-Bound								
High Temp.	-1.291	18.116	15.024	0.858	0.908	18.018	19.124	14.712
Low Temp.	-1.418	18.179	15.084	0.861	0.911	18.175	19.294	14.825
Wind Speed	-1.492	18.097	15.019	0.860	0.909	18.303	19.494	14.948
Cloud Cover	-0.676	18.833	15.900	0.901	0.987	22.253	23.282	17.927
Rain	-1.526	18.126	15.045	0.862	0.911	18.365	19.528	14.983
Thunderstorms	-1.731	18.147	15.061	0.854	0.905	18.144	18.990	14.690
Snow	-1.717	18.189	15.084	0.858	0.912	18.800	19.852	15.223
Precipitation	-1.494	18.090	15.013	0.860	0.909	18.275	19.417	14.882
Evaluation								
High Temp.	-1.295	18.116	15.024	0.858	0.907	18.020	19.155	14.709
Low Temp.	-1.409	18.156	15.064	0.860	0.910	18.157	19.268	14.812
Wind Speed	-1.304	18.180	15.084	0.864	0.913	18.362	19.604	15.073
Cloud Cover	-0.923	18.832	15.930	0.901	0.991	22.409	24.045	18.232
Rain	-1.754	18.144	15.066	0.863	0.912	18.518	19.693	15.141
Thunderstorms	-1.773	18.166	15.075	0.855	0.905	18.168	19.014	14.708
Snow	-2.157	18.322	15.203	0.865	0.919	19.257	20.433	15.687
Precipitation	-1.821	18.132	15.060	0.863	0.912	18.574	19.761	15.198

Table 4: Mean Demand Forecast Quality Measures by Included Exogenous Weather Variable

Bias (ME) across all models tended to be negative on average. This negative bias was amplified in models with various precipitation-based effects (for both *evaluation* and *upper-bound* models), and was lower in *4PL* models. Mean RMSE and MAE were also lowest in *4PL* models, and were lower in the *baseline* models than in models that included weather effects. A lone exception in the *upper-bound* models were models that included overall precipitation data, which had on average slightly lower RMSE and MAE. Scaled error (MASE) was on average slightly lower in both *evaluation* and *upper-bound* models that included thunderstorm data, but equal or slightly worse than the *baseline* models with all other weather effects. We could not calculate this metric for *4PL* models, as it depended on unavailable in-sample data. Relative measures (Theil's U) returned similar results, with thunderstorm data (and curiously predicted high temperature data) resulting in slightly lower Theil's U scores on average. As with scale-dependent metrics, *4PL* models had the lowest relative measures.

Percent errors represent the most interesting results, as *4PL* models that had dominated scale-dependent and relative measures were on average worse for percent errors. Mean MAPE, as measured over the first and second replenishment periods as well as over the weeklong planning period, was significantly higher in *4PL* models than in *baseline, upper-bound* or *evaluation* models. Further, inclusion of high and low temperature, as well as thunderstorm data resulted in lower average MAPE for both *evaluation* and *upper-bound* models. *Upper-bound* models with overall precipitation included also had lower MAPE. This curious result is likely due to the differences inherent in the metrics in use. Scale-dependent metrics penalize large deviations,

regardless of demand volume, and percent metrics experience inflation when demand is small. Therefore we can conclude that *evaluation, baseline,* and *upper-bound* models tend to be more accurate when demand is small, but when they miss, they miss larger than in the *4PL* models. Further, it seems that by including external weather effects in estimation, this effect is exacerbated.

In addition to the aggregate effects we observe between the forecast methods and included predictive weather factors, we also examined differences in demand forecast quality measures that were likely due to regional or product-related heterogeneities.

Table 5 and Table 6 display differences in forecast quality metrics among *evaluation* models based on NOAA region and menu category respectively.

NOAA Region	*ME*	*RMSE*	*MAE*	*MASE*	*Theil's U*	*MAPE 1*	*MAPE 2*	*Weekly MAPE*
Central (Ohio Valley)	-1.947	18.388	15.368	0.848	0.926	19.252	20.452	15.639
East North Central (Upper Midwest)	-3.371	19.814	16.506	0.850	0.885	20.113	18.179	15.039
Northeast	-2.540	18.446	15.410	0.892	0.965	22.224	24.810	18.723
South	-1.460	18.232	15.124	0.865	0.904	18.354	19.124	14.732
Southeast	-1.579	17.897	14.854	0.885	0.951	19.895	21.614	16.614
Southwest	-1.387	17.793	14.633	0.817	0.903	19.162	19.646	15.060
West	-0.449	18.268	15.158	0.862	0.896	16.678	17.576	13.565
West North Central (Northern Rockies and Plains)	-1.753	17.217	14.247	0.838	0.882	19.708	18.939	14.916

Table 5: Mean Demand Forecast Quality Measures by NOAA Region for *Evaluation* Models

Menu Category	*ME*	*RMSE*	*MAE*	*MASE*	*Theil's U*	*MAPE 1*	*MAPE 2*	*Weekly MAPE*
Add-on/Specialty	-0.317	9.739	8.060	0.859	0.849	13.701	13.973	10.727
Breakfast Entrée	-2.386	13.836	11.483	0.889	0.893	20.085	20.840	15.804
Breakfast Side	-5.445	32.840	27.100	0.909	0.916	20.597	20.962	16.014
Dessert/Shake	-0.836	10.088	8.300	0.731	0.875	26.408	27.531	20.629
Drinks	-2.242	16.522	13.721	0.876	0.909	15.675	16.706	12.956
Lunch/Dinner Entrée	-0.035	19.842	16.583	0.886	0.954	16.962	18.809	14.662
Lunch/Dinner Side	-2.420	27.614	23.035	0.893	0.956	18.039	19.159	14.777

Table 6: Mean Demand Forecast Quality Measures by Menu Category for *Evaluation* Models

Scale-dependent metrics tended to be worse in the Upper Midwest and Northeast regions, which could be a result of greater proportions of inclement winter weather in those areas. Severe weather can cause large misses in demand estimates, which are indicated by scale-dependent measures. The evaluation period spans mostly winter months, so these areas are more likely than Southern regions to experience inclement winter weather. Scaled, relative and percent measures tended higher primarily in the Northeast and Southeast.

Among menu categories, side items tended to perform worse among both scale-dependent and scaled metrics. Dinner items performed worst among dependent measures, and desserts were significantly worse than all others for percent errors. Significant heterogeneities exist between both NOAA regions and menu categories, and so the use of these indicators as controls is justified when evaluating the effect of weather variable incorporation on demand forecast quality measures.

We next conducted an analysis of variance for *upper-bound* and *evaluation* models that include all previously identified control variables. In the *upper-bound* models that include observed weather, we observe the asymptotic limit of the effect of weather forecast information on demand forecast quality measures. If the included weather information did not introduce additional variance, these are the results we would expect. Table 7 displays the results from inclusion of error-free weather variables.

	ME	***RMSE***	***MAE***	***MASE***	***Theil's U***	***MAPE 1***	***MAPE 2***	***Weekly MAPE***
Intercept	-2.204 ***	27.217 ***	22.836 ***	0.938 ***	0.893 ***	6.899 ***	8.005 ***	6.472 ***
Weather Variables								
High Temp.	0.223 ***			-0.002 *		-0.284 ***	-0.349 ***	-0.221 ***
Low Temp.	0.096 **				0.004 ***	-0.127 *	-0.179 **	-0.108 *
Wind Speed								
Cloud Cover	0.838 ***	0.738 ***	0.885 ***	0.040 ***	0.079 ***	3.953 ***	3.809 ***	2.995 ***
Rain				0.002 .	0.003 ***			
Thunderstorms				-0.002 .	-0.003 ***	-0.333 ***	-0.529 ***	-0.287 ***
Snow	0.171 ***			-0.003 *	-0.006 ***	-0.334 ***	-0.579 ***	-0.421 ***
Precipitation								
Covariates								
CV	1.878 ***	-23.829 ***	-20.388 ***	-0.213 ***	0.046 ***	32.473 ***	33.169 ***	24.254 ***
Station Distance		-5.47E-2 ***	-4.80E-2 ***	-2.29E-4 ***	-1.23E-4 ***		-1.81E-2 ***	-6.91E-3 ***
NOAA Region	***	***	***	***	***	***	***	***
Menu Category	***	***	***	***	***	***	***	***
Significance codes: ''= <0.001, '**'= <0.01, '*'= <0.05, '.'= <0.1***								

Table 7: Regression Effects of *Observed* Weather in Demand Forecasts on Forecast Quality Measures

Evaluation models represent a more realistic state for demand planners. Each weather prediction may have varying reliability depending on weather phenomena, range and proximity of a restaurant to a weather station. Each of these considerations makes the inclusion of predictive factors a more complicated and questionable decision. The results of models that include weather predictions in the estimation of demand forecast models are included in Table 8. A brief discussion of the implications of these findings on each hypothesis posed earlier follows, summarized in Table 9.

	ME	***RMSE***	***MAE***	***MASE***	***Theil's U***	***MAPE 1***	***MAPE 2***	***Weekly MAPE***
Intercept	-2.248 ***	27.219 ***	22.845 ***	0.939 ***	0.895 ***	6.814 ***	7.895 ***	6.379 ***
Weather Variables								
High Temp.	0.220 ***			-0.002 .		-0.282 ***	-0.318 ***	-0.224 ***
Low Temp.	0.106 ***				0.002 *	-0.144 **	-0.205 **	-0.120 **
Wind Speed	0.210 ***			0.004 ***	0.006 ***		0.131 *	0.141 ***
Cloud Cover	0.592 ***	0.737 ***	0.915 ***	0.041 ***	0.083 ***	4.109 ***	4.573 ***	3.299 ***
Rain	-0.240 ***			0.003 **	0.004 ***	0.220 ***	0.224 ***	0.211 ***
Thunderstorms					-0.002 .	-0.304 ***	-0.468 ***	-0.249 ***
Snow	-0.233 ***	0.174 *	0.118 .	0.004 **		0.109 .		
Precipitation	-0.307 ***			0.003 **	0.004 ***	0.274 ***	0.289 ***	0.267 ***
Covariates								
CV	2.035 ***	-23.854 ***	-20.425 ***	-0.216 ***	0.043 ***	32.833 ***	33.358 ***	24.492 ***
Station Distance	-2.64E-3 *	-5.41E-2 ***	-4.77E-2 ***	-2.29E-4 ***	-1.29E-4 ***	-6.73E-3 ***	-1.44E-2 ***	-6.17E-3 ***
NOAA Region	***	***	***	***	***	***	***	***
Menu Category	***	***	***	***	***	***	***	***
Significance codes: ''= <0.001, '**'= <0.01, '*'= <0.05, '.'= <0.1***								

Table 8: Regression Effects of *Predicted* Weather in Demand Forecasts on Forecast Quality Measures

By including high temperature in demand forecasts, we demonstrated significant reductions in negative bias, scaled and percent errors. However, there were no significant effects on scale-dependent or relative measures. These effects are consistent between *upper-bound* and *evaluation* models in both significance and relative effect size, indicating that this effect is robust to some degradation in forecast accuracy. This partially supports *H1a*, that inclusion of high temperature predictions improves demand forecast quality.

Inclusion of low temperature in demand forecasts produced similar results, if generally at a lower magnitude. Both *upper-bound* and *evaluation* models showed a decrease in negative bias, and significant reductions in percent error measures. However, there was no significant effect on scale-dependent or scaled measures, and a small but significant increase in Theil's U. This indicates partial support for *H1b*, that inclusion of low temperature predictions improves demand forecast quality. However, improvements tended to be in measures that face inflation from error coinciding with low demand. Theil's U, scaled by RMSE and therefore more sensitive to large errors regardless of coinciding demand level, saw a slight increase.

H1c was also partially supported, that forecast models incorporating short range predictions (one to three days) of temperature are more accurate than those incorporating medium range predictions (five to nine days). While it is true that error was higher in forecasts at a longer range, the effect of temperature predictions on percent error is greater at longer ranges. Table 4 shows that MAPE is higher for the second replenishment period under all *upper-bound* and

evaluation models, regardless of weather effect included. However, as shown in

Table 7 and

Table 8, the magnitude of error reduction through inclusion of temperature variables is greater at longer ranges. This means that the potential for error reduction through inclusion of external predictive factors is greater when the demand forecast is more uncertain, regardless of the uncertainty of the predictive factor. This contradicts the theory behind, if not the explicit statement of *H1c*.

Wind speed inclusion appears to have no effect on any measure of demand forecast quality among *upper-bound* models, and in *evaluation* models actually increases measures of scaled, relative and percent error. Despite also reducing negative bias in *evaluation* models, the overall effect of wind prediction mostly contradicts *H2a*, or that inclusion of wind predictions will increase demand forecast quality. *H2b* on the other hand, is supported. Less accurate medium range wind forecasts tend to degrade demand forecasts to a greater extent than short range wind forecasts.

Inclusion of cloud cover data in both *upper-bound* and *evaluation* models significantly increases scale-dependent, scaled, relative and percent error, while decreasing negative bias. As a result, *H3* is mostly not supported. It is worth noting, this measure also has no indicator of accuracy, so nothing definitive could be said about the effect of including perfect cloud cover prediction.

Given perfect prediction of rain, as in *upper-bound* models, the only significant effects are a slight increase in scaled and relative error. In *evaluation* models, predicted

rain significantly increases negative bias, scaled, relative and percent error. *H4a*, that inclusion of rain predictions in demand forecasts will improve accuracy, is not supported.

Forecasts including thunderstorm data, on the other hand, demonstrated significant improvements in scaled, relative and percent errors in *upper-bound* models, and in relative and percent errors in *evaluation* models. As a result, *H4b* is partially supported.

Inclusion of snow weather data resulted in significant reductions of negative bias and improvements in scaled, relative and percent errors in *upper-bound* models. However, these effects were reversed in *evaluation* models. Negative bias was increased, and scale-dependent, scaled and percent errors all showed some degree of increase. *H4c* was not supported.

Models that included overall precipitation data in *upper-bound* models showed no significant differences in demand forecast quality measures of any kind. In *evaluation* models, negative bias was exacerbated, while scaled, relative and percent errors all increased. *H4d* was therefore not supported.

H4d, the supposition that forecast models incorporating short range precipitation data (of all kinds) are more accurate than medium range is partially supported. While longer range demand forecasts each tended to be less accurate regardless of which weather forecast variable was included, the directional effect of including each variable was amplified at longer ranges. This means that predicted thunderstorm data decreased error to a greater extent at longer ranges. Conversely, predicted rain and overall precipitation inclusion increased demand forecast error slightly more at longer ranges.

	Variable	*Support*	*Explanation*
		Temperature	
H1a	High (^{o}F)	Partially Supported	Reduced negative bias, scaled and percent errors (supports), no other significant effects (does not support)
H1b	Low (^{o}F)	Partially Supported	Reduced negative bias and percent errors (supports), increased relative error, no other significant effects (does not support)
H1c	Medium Range	Partially Supported	Increased error at longer range (supports), error reduction greater at longer range (does not support)
		Wind	
H2a	Speed (mph)	Mostly Not Supported	Reduced negative bias (supports), Increased scaled, relative and percent error, no other significant effects (does not support)
H2b	Medium Range	Supported	Increased error at longer range and error amplification greater at longer range (supports)
		Cloud Cover	
H3	Percent	Mostly Not Supported	Reduced negative bias (supports), Increased scale-dependent, scaled, relative and percent error (does not support)
		Precipitation Related	
H4a	Rain Probability	Not Supported	Increased negative bias, scaled, relative and percent error, no other significant effects (does not support)
H4b	Thunderstorm Probability	Partially Supported	Reduced relative and percent error (supports), no other significant effects (does not support)
H4c	Snow Probability	Not Supported	Increased negative bias, scale-dependent, scaled and percent error, no other significant effects (does not support)
H4d	Total Precipitation Probability	Not Supported	Increased negative bias, scaled, relative and percent error, no other significant effects (does not support)
H4e	Medium Range (all kinds)	Partial Support	Increased error at longer range and error amplification greater at longer range for rain and total precipitation (supports), error reduction greater at longer range for thunderstorms and error amplification reduced at longer range for snow (does not support)

Table 9: Summary of Hypothesis Test Results

Discussion

Upper-bound models demonstrate the potential for improvement in demand forecast quality through inclusion of a number of external weather data under ideal conditions. In particular, future information on both high and low temperature and extreme weather such as thunderstorms and snow promised to improve a number of demand forecast quality measures, providing the weather data contained no errors. In *evaluation* models, this effect persisted for high temperature, low temperature and thunderstorm predictions, but reversed for predictions of snow. Despite potential improvements in both *upper-bound* and *evaluation* models, the inclusion of some weather variables degraded predictions, and all had disparate effects on the various measures of demand forecast quality. These mixed results we experienced demonstrate that incorporation of external variables do not have straightforward effects, and that the decision to include predictive indicators in an effort to improve demand forecasts depends on a number of factors.

Specification Errors

Estimating demand as we have depends on some key assumptions inherent to linear regression models. As Cohen et al. (2013) note, these include a correct specification of the form of the relationship between independent (in our case external weather indicators) and dependent variables (demand), that independent variables are correctly specified (are significant), and that the independent variables are measured without error. When including uncertain weather information in demand forecasts, it is likely that at least one of these assumptions is not strictly satisfied. However this does

not typically matter in practice, as models that perform better are chosen regardless of adherence to strict statistical orthodoxy (Hyndman and Athanasopoulos 2014).

However, ignoring these assumptions can introduce error in estimation and lead to over-fitting and misspecification. As SARIMAX models are estimated in two stages, the risks of estimation error can be compounded. Even if we assume independent variables are measured perfectly, as was the case in our *upper-bound* models, their relationship can be improperly specified or result in non-significant (or weakly significant) relationships. Such weak relationships often include some conflation of truly random error when estimated, which is then carried to the next stage of estimation assumed to be a genuine relationship. This has been shown to degrade prediction performance (Kolassa 2016a, Katsikopoulos and Syntetos 2016). We assumed linear relationships between weather factors and demand in the first stage, and estimated seasonal ARIMA terms from the residuals of that model. This can result in spurious explanation of random variance, and confound relationships that may exist in the second stage of estimation. Misspecification of this type will inevitably explain additional variance, whether or not the measured relationships are spurious, so we controlled for this by minimizing our automatically generated models on a measure (AICc) that includes a penalty for model complexity. Even so, overfitting can prove problematic with predictive models as evaluation transitions from in-sample to out-of-sample. Our results in *upper-bound* models for models including error free wind speed and overall precipitation data, which proved non-significant or even deleterious across all forecast quality metrics, may have been a result of such overfitting.

Confounding

Significant effects of the inclusion of error free weather predictors on demand forecast quality measures may have also been affected by confounding by product and regional heterogeneities. We did control for NOAA region and menu category, and the effects of these covariates were highly significant. However, errors in how these regions or menu categories are specified can have confounding effects for a model. For instance, boundary conditions between NOAA regions are likely similar in effect, but are assigned the mean effect for their respective regions. These discrete differences make estimation tractable, but can lead to distortion of effect estimation.

Weather Forecast Reliability

The varying reliabilities of weather factors may also have driven mixed results. This is evident in the degradation of demand forecast quality between *upper-bound* and *evaluation* models that included wind speed, cloud cover, rain, snow and overall precipitation data. Each weather variable had differing reliability in *evaluation* models, and therefore experienced different levels of degradation between *upper-bound* and *evaluation* models.

For *evaluation* models including wind speed, the degradation of the effect of predictive indicator inclusion on demand forecast quality measures indicates a possible combination of effects. Insignificant effects from these predictors observed in *upper-bound* models may have been from overfitting or misspecification. Including uncertainty in the predictors introduces noise to models that are already potentially misspecified, amplifying this effect.

The negative effect of including cloud cover is possibly a function of the available data for cloud cover. Though many prior studies found support for levels of sunlight significantly affecting demand (Johnston and Harrison 1980, Murray et al. 2010, Busse et al. 2012, Li et al. 2017, Steinker et al. 2016), we only had predicted cloud cover data available and thus no means of registering its relative accuracy. Our threshold of 30 km for the distance between a restaurant and weather station are consistent with expectations for reliable sunlight forecasts *on the aggregate* (Camargo and Hubbard 1999). However Camargo and Hubbard (1999) directly measured solar radiation in their study, whereas other studies measure binary sunny or non-sunny days (Li et al. 2017), hours of sunlight (Johnston and Harrison 1980, Parsons 2001, Murray et al. 2010), sunlight as an input to a composite weather measure (Steinker et al. 2016), or predicted percent of cloud cover (Busse et al. 2012) to ostensibly measure the same effect. These proxies may all have varying levels of reliability that corrupt their effect as previously reported and negatively impact their value to demand forecasters.

Distinct aspects of rain, thunderstorm, snow and overall precipitation prediction reliability may help explain our mixed results as well. Of the four precipitation-based weather predictors, only models including thunderstorm predictions did not degrade significantly in demand accuracy between *upper-bound* and *evaluation* models. Brier scores alone could not account for this difference, as both snow and thunderstorm predictions had low scores, and inclusion of uncertain snow forecasts in demand forecasts degraded forecast quality measures to a greater extent than did inclusion of uncertain thunderstorm forecasts. Thunderstorm predictions had the lowest PPV and highest

sensitivity of the four measures, indicating the lowest conservatism of the four precipitation-based weather predictors. This implies that demand forecast quality is not as adversely affected by false positives as it is by missed predictions. Demand managers may use this insight to their advantage when selecting weather forecast services or classification probability cutoffs for precipitation-based weather predictors.

Differential Effect between Demand Forecast Quality Measures

Variation in the effect of inclusion of exogenous weather predictors are also a function of the metric in use. None of the included weather predictors in either the *upper-bound* or *evaluation* models demonstrated an improvement in the scale-dependent-metrics RMSE or MAE. In *evaluation* models, weather predictors tended to only slightly improve and more likely degrade scaled and relative error metrics. For the included weather predictions that significantly improved demand forecast error, improvement came in the percent error metrics. These differences depend on the manner in which the SARIMAX models were estimated, and the relative penalties imposed by each demand forecast quality metric.

Predictive models were based on a minimization of AICc based on in-sample demand and observed weather. That primary improvement occurred in percent error metrics, that experience inflation in error coinciding with low demand, implies that inclusion of external weather predictors improves accuracy when demand is low. Increased error or insignificant effects on scale-dependent metrics imply that forecast responsiveness is higher and may lead to larger individual errors when demand is higher. Increases in scaled and relative errors imply an increase in responsiveness of demand

forecasts approaching that of a naïve forecast, with larger individual errors when demand is higher.

Limitations

Limitations of our research included deficiencies in the weather and POS demand samples, lack of insight into the *4PL* model parameterization and adjustments as a basis of comparison, and subjectivity regarding assignment to NOAA regions and menu categories.

One year of in-sample data allowed for a characterization of seasonal demand patterns for both demand and weather. However, we are limited to the assumption that conditions which drive changes in demand are stationary, which may not be the case given shifts in local and national tastes and stages of menu-item life cycle. Without more demand data, characterization of trends or shifts are more limited. Similarly, weather patterns can demonstrate anomalies from one year to the next, even on a national level. As noted in Hu and Skaggs (2009), though weather forecast reliability has increased in recent years, anomalies and associated reductions in medium-range weather forecast reliability are likely to increase in frequency as the effects of global climate change continue to manifest. Our out-sample evaluation data covers primarily winter months. Though we have demand history that would indicate effects during other seasons, we have no indication of model performance other than in the winter season. Additionally, our weather samples were drawn from a sensor network with sparse geographic coverage, resulting us having to compromise station relevance for sample size.

By virtue of the sample's nationwide scope, these findings are generalizable over a large number of circumstances. However, the results of this research are limited to a single (albeit industry leading) quick service restaurant firm.

As noted above, boundary conditions of NOAA region assignments may be a cause for additional effect distortion, and menu categorization is an admittedly subjective assignment. It is possible that a more exploratory clustering or principal components analysis approach could reveal a better of regional or product based grouping mechanism.

Managerial Implications

This research suggests that inclusion of short term predictive weather indicators for high temperature, low temperature and thunderstorms can significantly reduce demand forecast percent errors. It also indicates that the benefits of including weather predictions is greater as forecast horizon is increased, despite a decrease in weather forecast accuracy. These results apply to a wide range of geographic and product specific contexts. However, demand planners must take care when including other weather predictions, as they can have negative effects on demand forecast quality. The positive or negative effects of including predictive weather indicators in demand forecasts depends on factors such as weather phenomena forecast reliability, demand forecast error metric of interest, and heterogeneous effects that may exist between regions and products.

While such external predictive weather indicators are constantly improving, they currently have a reliable range that is on the boundary of being useful for most operational planning. As weather prediction technology and weather forecaster skill increases, managers can place more trust in these external indicators. However, even

reliable indicators should be treated with caution for the specification and confounding issues discussed above.

As each demand forecast quality measure responded differently to inclusion of additional information, managers should also carefully select which quality measures are important to their business. In our case study, inclusion of weather variables tended to make demand forecasts more responsive. When this improved measures of demand forecast quality, it occurred most in relative measures. This means that for businesses where large misses in high volume locations are relatively more expensive than a series of small misses in low volume situations, inclusion of weather in demand forecasts may not help operations. This may be the case if inventory costs are high and include high proportions of perishable goods. The opposite may be true for businesses where costs from low customer satisfaction are relatively more significant. Outlets facing high levels of competition may risk more from a stock-out than from overstocking. Therefore, the unequal improvement between measures of forecast quality require a manager to assess which measures are most relevant to their situation when including external weather predictions.

Future Research

The findings in this research provide support for a growing body of work relating observed weather phenomena to demand and predicted weather to demand forecast quality. We supported previous findings that inclusion of predicted temperature can significantly improve demand forecast accuracy (Nikolopoulos and Fildes 2013, Steinker et al. 2016), while motivating greater investigation into findings that were not supported,

relating predicted sunlight and precipitation to increased demand forecast accuracy (Steinker et al. 2016). The results of this work are limited to a select few forecast quality metrics. Future research must include more measures of quality, but also of overall value.

Some of the counterintuitive results also indicate other relevant information about the observed weather. Previous work estimating the effect of observed weather on demand includes studies that look specifically at how different the observed effect is from the average (Johnston and Harrison 1980, Tran 2016). We include no indication of the unusualness of a prediction or observed weather effect. Future research should extend these previous works to include some indication of the unusualness of a predicted weather effect on demand forecast accuracy.

We also do not include potential substitution indicators. Many of the restaurants in our sample share a common weather station from which their weather predictions are gathered. They could be close enough to serve as substitutes for each other's demand. The same may be true of similar competing restaurants. Future research should attempt to control for the confounding effects of substitution.

Travel intensity is another potential factor that could significantly interact with weather in its effect on demand. Restaurants that experience a significant portion of their demand from traveling customers, as may be the case in airports and along highways, may experience effects from weather unrelated to mood and psychology. Effects from delayed flights, closed roads, or other travel delays may confound or amplify weather effects and contribute to poor demand forecasts. Controlling for individual restaurant

characteristics could improve future efforts to incorporate predicted short term weather in demand forecasts.

Previous work in the economic value of uncertain information (Arrow 1965, Katz and Murphy 1997), especially research on cost-loss utility functions (Thompson and Brier 1955, Thompson 1962, Murphy 1977, Katz and Murphy 1990, Katz and Lazo 2011), can help guide future extensions to convert demand forecast quality improvements into calculations of expected value. This research included some discussion of differential effects of included exogenous weather forecast variable reliability on demand forecast quality. However, a more explicit treatment of the effects of accuracy, bias, sensitivity and specificity by predictive weather factor is warranted in order to fully quantify the effects on demand forecast quality, and eventually value.

Finally, this research estimated models based on linear and stationary effects of predictive weather indicators. Prior research suggests that many weather effects are nonlinear (Murray et al. 2010, Lazo et al. 2011, Bahng and Kincade 2012, Arunaj and Ahrens 2016, Tran 2016) and short to medium term weather forecasts will face increasing non-stationarity (Hu and Skaggs 2009), so future extensions ought to include more general estimation techniques.

Chapter 4: "A Systematic Literature Review and Typology of Factors that Bound Demand Forecast Accuracy"

Introduction

Demand forecasting is the bellwether that synchronizes the supply chain. Common thought regarding this critical input, among both practitioners and academics is that achieving more accurate forecasts is universally beneficial to the supply chain. Whether measured as the driver of safety stock, of variance in production processes, or bullwhip propagating throughout the supply chain, accuracy in forecasts goes a long way in ensuring efficient supply chain operations (Silver et al. 1998).

It is rare in academic literature for the somewhat counterintuitive question to be asked: "How good is good enough?". However, this is of central importance to the demand planner. We know that demand forecast accuracy is critical and that more is generally desirable in the supply chain, but what are the achievable limits? For that matter, are there instances where demand planners ought not even pursue the achievable? While a tremendous amount of academic literature investigates statistical, managerial, or technological means to improve demand forecast accuracy (Winklhofer et al. 1996, Fildes et al. 2008), few even concede the Pareto limits or costs to advances in precision (Yokum and Armstrong 1995). Current exploration also exists in vertically isolated channels, without an overarching frame to guide inquiry.

In this paper, we endeavor to identify the effects and contexts that have been explored in logistics and supply chain management literature that bound the feasible region for demand forecast accuracy. In doing so, we also identify areas that require further inquiry with the hopes of guiding future academic engagement. This search includes managerial, statistical, technological, and contextual facilitators and impediments for forecast accuracy. To accomplish this, we conducted a systematic review of the logistics and supply chain management literature, and identified structural topics regarding drivers of both achievable and desirable levels of forecast accuracy.

This manuscript is organized as follows. First, we define forecast accuracy and discuss general logistics and supply chain management research topics that may affect, or be affected by, accuracy. Next, we detail our methodology of systematic literature review, emphasizing transparency and replicability. Then, we discuss extant findings in academic literature within identified logistics and supply chain management research themes, describing managerial implications and suggesting future research areas for each theme. Finally, we discuss limitations to this study, and draw conclusions regarding the identified structural themes of demand forecast accuracy.

Defining Accuracy in Logistics and Supply Chain Management Research

Makridakis and Wheelwright (1989) refer to forecast accuracy as the "goodness of fit" of some predictive model. This obviously has different connotations for each user, application, and modeling approach. Murphy (1993) defines forecast accuracy as the correspondence of individual observations and predictions. Though he was referring to weather forecasts, this is equally applicable to demand forecasts. Box and Jenkins (1976)

express accuracy as the probability limits of a forecast such that some proportion of realized values fall within it. This takes a statistical viewpoint, and implies a distribution rather than point forecast. Armstrong (2002) defines forecast accuracy only in its relation to forecast error, with error being the difference between a forecasted and observed values. Wooldridge (2015), like Armstrong, defines accuracy in relation to error, but as the additive inverse of Armstrong's error. Both Armstrong's and Wooldridge's definitions recognize that it is often easier and more useful to measure when a forecast is wrong than when it is right. Silver et al. (1998) defines forecast accuracy simply as a surrogate for overall production/inventory system performance. This definition lacks precision, as many factors affect system performance beyond demand forecasting accuracy. Makridakis and Hibon (2000) and Hyndman and Koehler (2006) note multiple measures of forecast accuracy that each prioritize different aspects of error, and conclude no one measure fully reflects accuracy. While neither work explicitly defines accuracy, they do present strengths, weaknesses and applications for numerous proposed measures of forecast error (as proxies for accuracy).

Among the various definitions, there is a general consensus that forecast accuracy is a measure of absolute closeness of a prediction to observed conditions and a complement to error. We choose to adopt this more general definition, as many slight variations exist on how accuracy and/or error are measured and applied.

While this study focuses on forecast accuracy, many forecast evaluation constructs are derived from, or are used as proxies for accuracy. Terms such as precision, bias, deviation, error, quality, performance, consistency, and reliability all may overlap

the concept of accuracy, but have definitions that vary depending on source and usage. We explore these concepts, in addition to the concept of accuracy, as they are discussed in a wide range of supply chain management and logistics contexts that may help to categorize our understanding of the effect of forecast accuracy.

Basic Overview of Systematic Literature Review

To address the question of "How good is good enough?", we conducted a structured literature review of research that has been done with respect to the bounds of forecast accuracy. A structured or systematic review involves searching, selecting, appraising, interpreting and summarizing of data from original studies (Crowther and Cook 2007), and serves as the highest level in a hierarchy of evidence (Tranfield et al. 2003). In essence, it is the most complete manner of characterizing the state of knowledge on a given subject, and a methodological advance over the narrative literature review used more often in management research (Denyer and Neely 2004). Imposing structure in a review helps make the process of knowledge collection and generation transparent and replicable, while reducing the impact of researcher bias (Durach et al. 2017).

Following the advice of Tranfield et al. (2003), and expounded on by Durach et al. (2017), we divide our literature review into several steps. The first is to define a research question, as we have done above. All subsequent steps ought to be guided by this original research question. Second, we determine the criteria for inclusion in a review. This includes criteria to ensure relevance to the research question, but also to ensure quality of source and tractability of the review itself. Third, we collect potentially

relevant research for review. We conduct this in multiple rounds. Based first on a keyword search in a research database, and then on review of the cited and citing literature of studies we have found to be relevant; a process we call "cascading". Fourth, we apply the inclusion criteria on the collected article sample. For each article returned in the keyword search, this is done separately based on a review of the abstract, and then on the full article for those abstracts that were deemed relevant. The same process is conducted for cascaded articles considered for inclusion. Fifth, we synthesize the relevant literature sample and develop themes around our research question. Sixth and finally, we present the results of our search.

Criteria for Inclusion in the Literature Review

Topics for the Literature Search

Determining criteria for article inclusion began with the development of several general topics and contexts that may have an effect on the upper and lower bounds of demand forecast accuracy. These were identified through discussions within the author team, and an initial search of the logistics and supply chain management literature for forecast accuracy. General topics initially emerged as factors that may drive or inhibit levels of forecast accuracy. These could be statistical, such as time series variability limits to prediction or model over-fitting. Topics could be managerial, such as the cost of gathering or processing information or the propensity of agents to distort or withhold information. These could also be technological, such as advances in information sharing technology or introduction of novel forecasting techniques. Below are the general topics that emerged from our initial search, and the guide for our structured literature review.

The complete list of Boolean search terms can be found in Appendix A: Boolean Keywords for EBSCO Business Source Complete.

Bullwhip Effect: Perhaps most prominent in the supply chain management literature is the demand signal propagation effect originally known as industrial dynamics (Forester 1958), but now more commonly referred to as the bullwhip effect (Lee et al. 1997). Forecast accuracy, and by extension all of the related constructs we listed, have been observed to affect the nature and amplify the magnitude of the bullwhip effect. Forecast error amplification can have significant negative cost effects within and between firms of a supply chain, and it is useful to identify the importance of the accuracy of the original demand signal relative to other factors contributing to the bullwhip effect. We therefore included "bullwhip" and several variants in our keyword search.

Cost Tradeoffs: Accuracy improvements almost always come at some additional cost. These costs must be balanced against the potential benefits derived from incremental improvements. To capture this dimension, we included (primarily) empirical research that strove to quantify either the incremental cost of forecast accuracy improvements, or the costs borne from inaction. For each specific circumstance, we preferred works that tried to calculate both. We included the term "tradeoff", and several variants as identified in two such works (Metters 1997, Sanders and Graman 2009).

Aggregation or Vertical Hierarchical Level: The hierarchical level at which the forecast is generated, and the level for which it is intended significantly affect the impact of forecast errors. While greater levels of aggregation tend to wash out noise and result in higher levels of accuracy, the resultant forecasts also eliminate much of the detectable

signal. This makes more aggregated forecasts more useful for strategic purposes, but of limited utility for operations. We include variations of "aggregation" defined in Clemen (1989) and Zotteri et al. (2005) to select research that empirically defines the contextual utility of various levels and types of forecast aggregation.

Impact of Manual Adjustments: Manual adjustments are often a necessity with the reality of imperfect methods, insufficient information, or unquantifiable demand factors. Furthermore, despite significant advances in both technology and methods to develop reliable statistical forecasts, a large proportion of businesses rely very little or not at all on quantitative forecasting methods. We include the term "judgement" and many variations identified in Sanders and Ritzman (1995, 2004b) in order to explore the significance of manual forecast adjustments on demand forecast error.

Impact of Information Sharing: A significant portion of research is dedicated to the cost and efficacy of various types of information sharing in the supply chain. There are many reasons to share information in the supply chain, but we are interested specifically in the value generated in demand planning as outlined in Lee et al. (2000) and Cachon and Fisher (2000). While obviously linked to the mitigation of the bullwhip effect, it constitutes multiple separate streams of research and is but one of many means of reducing the impact of forecast error signal amplification. For this reason, we include "information sharing" along with various derivatives as a keyword search term.

Supply Chain Collaboration and Integration: Related to the idea of information sharing, some supply chain partners go further and jointly forecast demand to smooth supply chain operations. We differentiate collaboration from information sharing, and

include it on a continuum of integration between supply chain partners that in the extreme results in vertical integration (Bowersox et al. 2013). We consider both vertical and horizontal collaboration, between complementary and also competing members of the supply chain. We included various instantiations of "collaboration" and "integration" as keywords to identify work that discusses the forecast accuracy ramifications of different arrangements.

Choice of Metric or Measure: Though we define accuracy above, there are still many ways to measure the closeness of a prediction to an observed condition. We include many ways of referring to this closeness in our search terms above, but we needed additional search criteria to identify the implications of specific types of measures in use. Answering the question "How good is good enough?" can have different answers depending on the accuracy metric in use. Makridakis and Hibon (2000) and Hyndman and Koehler (2006) served as a starting point to describe the contextual effects of the choice of error metric on desired (or achievable) forecast accuracy. We therefore include the search words "measure" and "metric".

Overfitting and Misspecification: One noteworthy area for forecasting research focuses on the difference between verification and validation. Forecast models typically are fit on available data, and extrapolated forward. Modelers verify that their models fit the available data, and the model is only validated when the extrapolation is sufficiently close to the observed outcomes. A bedeviling reality for the modeler is that models that fit very well to the past often do quite poorly in predicting the future (Armstrong 2002,

Kuhn and Johnson 2013). We include "over-fit" and its variations to capture the research that defines the role of model misspecification on forecast accuracy.

Lead Time and Inventory Policy: With regards to inventory costs, demand forecast inaccuracy likely has greater significance under conditions of longer lead time and when paired with specific inventory policies (Silver et al. 1998, Fildes and Kingsman 2011). Some combinations of forecast inaccuracy and replenishment policy may also significantly drive up transportation, handling, warehousing, ordering, and expiration or obsolescence costs. As a result, we include terms related to "lead time" and "inventory" in our search.

Forecast Horizon: As forecast horizons increase, demand forecast accuracy diminishes. It is useful, however to forecast to various lengths for different operational and strategic purposes. We include "horizon" and multiple variants as search keywords to include research that considers the need for accuracy at various horizons.

Product Inimitability and Supply Chain Flexibility: As described in Rajaram and Tang (2001), forecast accuracy may be relatively more important for supply chains in which there are few or no substitutes for a product or component. Conversely, accuracy may be substituted for with a more agile or flexible supply chain Christopher (2000), Chopra and Sohdi (2004). As such, we include variants of "substitute", "flexible" and "agile" in our keyword search.

Tolerance, Resilience and Point Forecasts: Weiland and Wallenburg (2012) describe a situation related to inimitability. Some supply chains are at greater risk, or conversely have a greater tolerance for ambiguity. Demand forecasting in these cases

often includes not only extrapolating past events, but quantifying the probability of demand factors as well. Forecasting for these purposes often requires more than just a point forecast, particularly when the uncertain event is binomial in nature. We include terms like "robust", "resilient", and "point" to capture this type of research.

Variability and Forecastability: The final set of search terms have to do with achievable limits of accuracy. Increased demand variability, nonstationary demand factors, lumpiness and intermittence all have been found to significantly limit the possible forecast accuracy (Syntetos et al. 2005, Gardner Jr. and Acar 2016). To capture research that systematically defines these conditions in the context of logistics and supply chains, we include terms such as "variability" and "uncertainty".

Scoping the Literature Search

Upon establishing the general search topics, our next step in developing criteria for inclusion was to find ways to limit a sample so it could be tractably vetted for quality. Initial searches in various databases returned thousands of articles, some from dubious sources and of questionable relevance. Several means of limiting the sample of a literature review have been suggested, including limiting the date range (Grimm et al. 2015), journal selection (Carter and Easton 2011), database selection (Winter and Knemeyer 2013), and citation count (Tate et al. 2012, Ellram and Cooper 2014).

To ensure the quality of the sources (Crowther and Cook 2007), and to limit the amount of unverified and un-vetted material we have to analyze (Eksoz et al. 2014), we bound our search to be peer-reviewed work that has a focus on logistics or supply chain management topics. In this way, we were able to define what the common academic

understanding of "How good is good enough?" was among logistics and supply chain management researchers over a wide array of contexts.

Two previous systematic literature reviews of logistics and supply chain management literature utilized Google Scholar, and scoped their search based on a prescribed number of top cited results from their keyword search (Tate et al. 2012, Ellram and Cooper 2014). We found three primary issues with this approach. First is that Google Scholar results tended to include a tremendous amount of noise. That is, search results that were not from peer reviewed sources, and citation counts that included non-peer reviewed citations. The second problem is a bias toward older articles, as citation counts (a heavily weighted criterion for relevance) for recent articles tended to be quite low. Finally, this method undercuts the "completeness" of a systematic search. By selecting an arbitrary number of highly cited or highly relevant articles, many practically relevant works are omitted.

We instead decided to limit our search to journals that are widely recognized as purveyors of high quality logistics and supply chain management research. We utilized the EBSCO Business Source Complete database, as this provided the most comprehensive collection of journals (and articles within journals) based on an initial keyword search in multiple databases.

In their systematic review of sustainable supply chain management research, Carter and Easton (2011) included a list of only seven journals as containing relevant research. We found this to be too restrictive, and included also those journals from the Supply Chain Management Journal list. This list, endorsed by over 300 university

professors conducting logistics and supply chain management research, expanded our total number of journals to 12. After consulting a leading logistics and supply chain management researcher who specializes in forecasting research, we expanded this list further to ten additional journals that included the highest number of articles that met our keyword search criteria, based on an unrestricted search.

The journals listed in Table 10 were included in the initial keyword search, which we divide into journals where logistics or supply chain management is their primary focus, and those that include relevant research, but have a primary focus elsewhere (such as operations research or general management topics):

Logistics and Supply Chain Journals	***Non-Logistics and Supply Chain Specific Journals***
Manufacturing and Service Operations Management	*Computers and Industrial Engineering*
International Journal of Logistics Management	*Decision Sciences*
International Journal of Physical Distribution and Logistics Management	*Decision Support Systems*
Journal of Business Logistics	*European Journal of Operational Research*
Journal of Operations Management	*Foresight: The International Journal of Applied Forecasting*
Journal of Supply Chain Management	*International Journal of Forecasting*
Production and Operations Management	*International Journal of Production Economics*
Supply Chain Management: An International Journal	*International Journal of Production Research*
Transportation Journal	*Journal of the Operational Research Society*
Transportation Research: Part E	*Management Science*
	Omega: The International Journal of Management Science
	Operations Research

Table 10: Journals Included in Initial Keyword Search

We opted not to limit our search based on date range, as we found no basis for deeming older work to be non-relevant, and were able to scope a tractable sample with existing limiters.

Individual Article Criteria for the Literature Search

Once the scope of the search is established, each article must be reviewed for relevance to the research question. Keyword searches often return false positives if search terms are too general, or omit relevant work if search terms are too specific. In response, we erred on the side of generality in the keyword search and subject each

article (in both abstract and full article reviews) to the following criteria to ensure it contributes to our intended research question. Each article:

1. Must help answer the question: "How good is good enough?"
2. Must at a minimum discuss the implications of differing levels of forecast accuracy, and this should be a main focus of the paper.
3. Should explore unique (and in combination, comprehensive) contexts that may affect requirements for accuracy that differ from simply "more accuracy is better".
4. Cannot simply assume greater accuracy is a sufficient goal across all contexts, must be application-specific, and agnostic to forecasting method (assumes most appropriate method used, and all reasonable control actions within the firm are exercised).
5. Must deal with demand forecasts (or those directly relating to the downstream flow of goods or services in a supply chain). This excludes economic, financial, demographic, managerial, technological, supply (or yield), or returns (though permits closed-loop systems) forecasts.
6. If forecast accuracy is not the main focus, then there must be some unexpected application or context-based insight regarding accuracy

Collecting Relevant Research and Applying Search Criteria

The first-round search returned 180 articles that met our keyword and journal restrictions. We reviewed the abstract for each, applying the six criteria for individual article inclusion, and categorized them on a five-point scale ranging from “clearly does not meet criteria for inclusion based on abstract, and therefore removed” to “clearly

meets criteria for inclusion based on abstract, and therefore merits full review of article". Based on this scale, all but the lowest category was retained for a full review, resulting in 130 carried forward. We then applied the same criteria for individual inclusion reviewing the full article. We found 107 articles met our criteria for inclusion after this second round.

From these 107 remaining articles, we began our cascading process. We reviewed all article titles cited in the bibliography and all citing articles identified in Google Scholar, using the individual article inclusion criteria. If a cited or citing article appeared to merit inclusion, and it was published in one of our previously identified journals, then it was carried forward for a review of its abstract. If it appeared in an alternate journal title, then the keyword search in Business Source Complete was expanded to include these journals. The result was an expansion to 191 articles, reduced to 181 articles after reviewing first abstracts, then full articles. Table 11 provides an overview of the cascading and vetting process.

Journal	*Initial Vetted Search*	*After Cascading*
International Journal of Production Economics	18	23
International Journal of Production Research	9	16
Journal of Operations Management	15	15
International Journal of Forecasting	5	15
Foresight: The International Journal of Applied Forecasting	5	14
Production Planning and Control	NA	12
Production and Operations Management	10	11
Journal of Business Logistics	9	10
Decision Sciences	2	7
Omega: The International Journal of Management Science	4	6
European Journal of Operational Research	2	6
Supply Chain Management: An International Journal	5	5
Computers and Industrial Engineering	4	5
Management Science	3	5
Manufacturing and Service Operations Management	3	5
Journal of the Operational Research Society	3	4
Operations Research	1	4
International Journal of Physical Distribution and Logistics Management	3	3
The International Journal of Logistics Management	3	3
Journal of Business Forecasting Methods and Systems	NA	3
Journal of Forecasting	NA	3
Naval Research Logistics	NA	3
Decision Support Systems	1	1
Journal of Supply Chain Management	1	1
Transportation Research: Part E	1	1
Transportation Journal	0	0
Total	***107***	***181***

Table 11: Keyword Search Results in Initial and Cascaded Search

Overview of Thematic Findings

As we reviewed our sample of literature, we identified primary and alternate themes (from the 13 general topics included in our search terms) that each article

included in their analysis. We noted the type of analysis conducted, the error measurement each work included (if any), and the supply chain entity(s) that served as their unit of analysis. Summary statistics for these characteristics can be found in Table 12, Table 13, and Table 14.

Type of Analysis	***No. of Articles***
Simulation, Non-Empirical	46
Analytic, Non-Empirical	35
Statistical Forecast Model, Empirical Validation	28
Thought Leadership	22
Case Study	21
Analytic, Empirical Validation	14
Regression Analysis of Forecast Accuracy	11
Survey	8
Simulation, Empirical Validation	7
Literature Review	6
Statistical Forecast Model, Non-Empirical	4

Table 12: Types of Analysis in Article Sample

Kolassa (2016b) Classification	*Accuracy Measure*	*No. of Articles*
Bias	Cumulative	6
	Mean (Median)	23
	Scaled Mean	2
Absolute Error	Mean (Median)	24
	Mean (Median) Relative	8
	Geometric Mean	2
	Geometric Mean Relative	3
	Mean (Median) Scaled	5
Absolute Percent	Cumulative	3
	Mean (includes minor corrections)	42
	Median (includes Relative)	5
	Symmetric Mean (Median)	12
	Weighted Mean	4
Percent Error	Mean (includes Symmetric)	10
Percent Squared Error	Mean Squared	1
	Root Mean (Median) Squared	1
Squared	Sum (includes Root Mean)	2
	Mean (includes minor corrections)	27
	Root Mean	8
	Relative Mean	2
	Relative Root Mean (Theil's U)	1
	Relative Geometric Root Mean	2
	Geometric Root Mean	4
	Coefficient of Determination	2
Scaled	Univariate Normal	40
	Univariate Non-Normal/Bayesian Updated	9
	Multivariate	5
	Coefficient of Variation	5
Functional	Loss Function Variants	5
	Error Implication Statistics	3
	Nonparametric (Ordered or Percent Better/Best)	3
NA	Self-Reported	4
	None	38

Table 13: Forecast Accuracy Measures Explored in Article Sample

Supply Chain Entity Studied	*No. of Articles*	*Supply Chain Echelons*
General Demand Forecasting	59	One-Tier
Distributor-Retailer	27	Two-Tier
Manufacturer-Retailer	23	
Supplier-Manufacturer	10	
Manufacturer-Multiple Channels	2	
Manufacturer-Distributor	1	
Supplier-Distributor	1	
Manufacturer-Distributor-Retailer	6	Three-Tier
Supplier-Distributor-Retailer	1	
Supplier- Manufacturer -Retailer	1	
Supplier-Manufacturer-Distributor-Retailer	3	Four-Tier
Production and Inventory Control	47	NA
Residual Demand	10	

Table 14: Focal Supply Chain Entities in Article Sample

We also found that we needed to edit our themes slightly to better reflect the extant literature and form a typology of drivers of "How good is good enough?". Themes can be divided generally into technical drivers of forecast accuracy bounds, and managerial drivers to accuracy bounds. Table 15 includes the six edited themes that we categorized as technical drivers and seven edited themes that fell under managerial drivers, and the numbers of articles where each theme was either a primary or an alternate focus. We then describe the state of logistics and supply chain management research regarding each theme, state what that means for our research question, and propose directions for future inquiry in academic research.

It is clear that many of the works reviewed focused on multiple dimensions we identified as being important to the question of "How good is good enough?". The

classification of each of these works as contributing to each dimension, while systematic, is ultimately subjective. Some research included merely a cursory and tangential discussion of the implications of accuracy. Some of the classifications could be viewed as equally contributing to another of the dimensions we set out. For instance, a study describing the bullwhip implications of forecast accuracy almost certainly include discussions of information sharing, the most prominent ameliorative tactic to reduce the impact of supply chain demand distortion. In that vein, a discussion of information sharing most certainly involves some focus on the degree of supply chain integration. This is then related to issues of hierarchy, aggregation, and the inevitable cost tradeoffs that occur within and between principals of a complex supply chain. Our judgement is based on the subjective set of criteria set out above, and a qualitative assessment of which dimension featured most prominently in a work.

Technical Drivers of Forecast Accuracy Bounds	***No. of Articles***	***Managerial Drivers of Forecast Accuracy Bounds***	***No. of Articles***
Forecastability	19	Error Amplification	8
Horizon	8	Cost Tradeoffs	43
Overfitting and Misspecification	4	Supply Chain Integration	34
Tradeoffs of Metrics	31	Supply Chain Flexibility	9
Level of Aggregation and Hierarchy	21	Manual Adjustments	11
Data Quality and Availability	12	Risk	24
		Inventory and Control Policy	34

Table 15: Technical and Managerial Drivers of Forecast Accuracy Bounds

Technical Drivers of Forecast Accuracy Bounds

While many elements of demand forecasting are under the control of managers or demand planners, technical drivers are either only partially endogenous or completely exogenous. These themes primarily drive achievable demand forecast accuracy. Literature findings, implications for "good enough", and proposed future research is discussed below for each technical driver of forecast accuracy bounds, summarized in Table 16.

Forecastability

Forecastability refers to the inherent limits to generating accurate future predictions from a particular demand pattern. In their survey of Midwestern businesspeople, Mentzer and Cox (1984) describe this concept of forecastability as a distinction between potential and achieved forecast accuracy. The reviewed literature seems to split between statistical and situationally-based inherent limits to forecastability. Statistical limits relate to characteristics the time series itself that make that pattern difficult to project. Situational limits refers to contexts likely to present these statistical limitations. Mentzer and Cox (1984) note that most work in demand forecasting focuses on methods to improve accuracy given an isolated context rather than a general classification of what is possible, much less the situational characteristics that may actually be under a manager's control. Despite the intense focus on methods, the discussion of what is forecastable remains underdeveloped. This is in part due to the concept being a moving target.

Statistical forecastability depends on the current state of constantly evolving forecasting methods and technology, the ability to gather data (covered more under a separate theme), and numerous data characteristics such as times series trend, seasonality, intermittence (Willemain et al. 2004, Hatzakis et al. 2010), (non)stationarity (Hosoda and Disney 2006, Gaur et al. 2007, Pearson et al. 2010, Thiel et al. 2014) and variability in both demand and supply (Lorentz et al. 2007, Fildes et al. 2009, Davis et al. 2016). As Fildes et al. (2008) suggest, improvement in statistical forecasting methods is not nearly as important as improving methods to match (particularly algorithmically) forecasting methods to circumstance. Enhanced statistical sophistication has not led to universal improvements in forecastability even though there is possibility for improvement. Additionally the focus of our review was not a technical discussion of forecasting methods, but instead of demand forecasting applications. As such, we include only data characteristics in our theme of forecastability and exclude a discussion of methods.

Situational forecastability is simply an examination of contexts in which the aforementioned statistical characteristics are likely. High demand variability can occur in any industry, and can be exacerbated by lead time or production instability, supply uncertainty, and the bullwhip effect between echelons in the supply chain. Davis et al. (2016) demonstrate the effect of supply uncertainty on forecastability by modeling food bank operations, though this can apply to other contexts such as nonprofits, agricultural supply chains, and extended supply chains with sourcing in unstable or poorly regulated areas. Lorentz et al. (2007) demonstrate the effect from lead time uncertainty in unreliable Eastern European logistics infrastructure. Intermittence is also notoriously

difficult to forecast owing to both variability in demand size and pattern (Syntetos and Boylan 2001), and has been studied in such contexts as residual demand for service parts (Willemain et al. 2004) and service demand (Hatzakis et al. 2010). Non-stationarity is observed in instances of rapidly changing demand conditions, and can be the result of a number of factors including firm (or product) age, level of market competition, industry, forecast horizon, level of aggregation, and market (or regulatory) maturity (Winklhofer 1996). Thiel et al. (2014) explore this in the effects of a health scare in the French poultry industry. This involves drastic and unpredictable reductions of demand in a tainted product, rapid increases in alternatives, uncertain supply, and unpredictable regulatory intervention. Short lived products with highly elastic consumer responses, like those found in the fashion retail (Pearson et al. 2010) and consumer technology (Gaur et al. 2007) industries also contribute heavily to non-stationarity. Even relatively stable and predictable point of sale demand can have non-stationarity introduced at higher echelons via bullwhip demand distortion (Hosoda and Disney 2006).

What these studies on statistical and situational forecastability indicate for practitioners is that some demand patterns commonly found in the situations described above will necessarily have lower achievable accuracy, meaning "good enough" is lower in these contexts. This also implies that alternate strategies to investing in more sophisticated forecasts will bear more fruit. Fildes et al. (2009) suggests that in situations of high variance (though not low), manual adjustments to forecasts are effective. Babai et al. (2014) suggest introducing bias known to be incorrect results in better operational outcomes than being able to correctly forecast intermittent demand complicated by such

factors as obsolescence. Chen and Blue (2010) suggest aggregating negatively correlated demands to offset low forecastability. Kurtuluş et al. (2012) suggest greater operational integration between supply chain echelons to limit variance inflation from bullwhip, but concede that this also likely has Pareto returns. Morlidge (2014b) suggests adjusting the manner in which forecastability is measured to prioritize how it is addressed. The most common measure associated with forecastability is the coefficient of variation (CV), or the ratio of standard deviation to mean in a time series. However, this is merely an indicator of dispersion which happens to correlate with poor forecast accuracy in most (but not all) cases (Morlidge 2013). Wallstrom and Segerstedt (2010) suggest also including indications of intermittence as well as dispersion, and Morlidge (2013, 2014 a,b) proposes measures like relative absolute error (RAE) as an alternative. This measure, which compares error in forecasts to error in a naïve forecast more appropriately demonstrates the potential forecastability relative to a naïve approach. He analytically demonstrates a theoretical lower bound of 0.7 and empirically finds a practical lower bound of 0.5 across numerous demand signals. Using this metric and some indicator of relative value contribution for each product, forecasters can better target limited funds to improve on low forecastable items.

Much academic work focuses on improving accuracy for demand predictions in the low forecastable situations listed above, and ought to continue in the future. However future research on forecastability should also examine interactions between these situations, and alternative means of both measuring and avoiding low forecastability.

Horizon

It should come as no surprise that previous logistics and supply chain management research has indicated that forecast accuracy degrades at higher forecast horizons, demonstrated in self response among demand planners (Mentzer and Cox 1984), and in simulations of multi-echelon production systems as lead time increases (Huang et al. 2011).

However lower accuracy does not necessarily mean they are not useful, as some research indicates the necessity of lower accuracy, longer range forecasts for point of sale demand. Inaccurate forecasts over longer horizons are necessary in products with long development lead times and short product life cycles (Fisher and Raman 1996, Li et al. 2015), and shortening lead time or increasing responsiveness to early indications of lead time inaccuracy promise greater returns than attempting to improve accuracy at those ranges. Clark (2005) and Amornpetchkul et al. (2015) show that inaccurate longer range forecasts also work as coarse indicators to suppliers, permitting production smoothing and more effective capacity planning. Miyoaka and Hausman (2004) show that the use of "stale" forecasts (longer horizon from previous period) for setting downstream base stock levels can help better align stock policy to production as forecasts are updated. Such examples of using inaccurate long range forecasts include benefits that are unequally shared and costs unequally borne (discussed further under the cost tradeoffs theme), often requiring built-in incentives or penalties in contracts.

The accuracy of longer range forecasts is also affected by the intended purpose of the forecast. Though disaggregated demand forecasts degrade at longer horizons,

temporal aggregation for long term operations, capacity, marketing or strategic planning tends to dominate the negative effect of forecast degradation due to horizon. Willemain et al. (2004) found certain methods for estimating intermittent demand (based on temporal aggregation) actually performed better with longer horizons as a function of the central limit theorem. This tradeoff between the accuracy benefit of aggregation with the deleterious effects of increased horizon is explored by Zhao et al. (2002b) in a simulation of a two-tier soft drink bottling production system. As forecast error increases, however, the benefits to suppliers of early order commitment decreases.

Increased horizon for disaggregated demand forecasts tends to lower achievable accuracy, and so indicates "good enough" is lower for situation where longer forecasts are warranted. Practitioners can offset this lower achievable accuracy by implementing postponement, electronic data interchange with agile production, and expediting when necessary (Li et al. 2015). Such investments, which can also be undertaken for short-term operational forecasts, promise greater overall cost savings with longer horizon forecasts (Rafuse 1995). However, these responses require a full understanding of both the incremental cost of inaccuracy and the proportion of error attributable to forecast horizon. This changes as the utility of the forecast is adjusted to longer range operational or non-operational purposes. Due primarily to temporal aggregation, "good enough" increases when longer range forecasts are combined for other planning purposes.

Future work regarding the effects of forecast horizon on demand forecast accuracy needs to include comparisons of the relative effect of horizon across industries, levels of hierarchy and various other contexts. It would also be useful to generate non-

linear models to characterize the degradation effect of increasing horizon on demand forecast accuracy.

Overfitting and Misspecification

This topic was somewhat less explored in the logistics and supply chain management literature, as it has a greater focus on pure statistical precision. Overfitting revolves around the tradeoff that exists between model complexity and precision, with regards to a training sample. As Kolassa (2016a) and Katsikopoulos and Syntetos (2016) describe in their discussions of tradeoffs between bias and variance, models with higher complexity are attractive as they tend to minimize bias and often "pass for intelligence". The problem is that unnecessary complexity attempts to explain variance that may be true random error, particularly when the underlying signal is weak. Katsikopoulos and Syntetos (2016) found more complex models increased out-of-sample forecast error by an average of 27%, and Kolassa (2016a) often found misspecified (biased) models outperforming correctly specified models. Even forecast model complexity increases that have promised improvements tend to be based on extremely small validation samples, and generalized claims from highly complex models should be viewed with skepticism pending further validation (Goodwin 2011). Willemain et al. (2004) demonstrate overfitting as a particular problem in intermittent derived demand situations. Empirically derived distributions for spare parts demand were found to be unreliable in predicting future demand when demand was non-stationary across 28,000 products. Their proposed alternative to accuracy was aggregation via cumulative lead time demand.

This implies that "good enough" is often lower than what is achievable. This is a more pernicious problem for demand planners, as overfitting is often a subjective assessment a-priori. That is, until the predictions are proven false, there is limited knowledge of their quality. The surest way to protect against this is to follow good modeling practices of including validation or holdout samples when fitting statistical models, and to employ demand forecasters that have a fundamental understanding of what is inevitably being optimized in the statistical models employed. As overfitting is likely amplified when demand is non-stationary, it is also useful to provide manual adjustments where possible to account for model deficiencies where it is known that demand pattern environmental conditions will change.

Future research on the effect of overfitting on demand forecast accuracy should focus on providing additional guidance for practitioners to detect overfitting. More work also needs to be done to describe the effect of the size and quality of the training sample on overfitting or misspecifying prediction models.

Tradeoffs of Metrics

Each metric of accuracy rewards and penalizes different aspects of closeness or difference of a prediction from reality. It is critical to match the measure appropriately to the circumstances. Unfortunately, in many cases, the circumstance is a demand planning function with little expertise that utilizes a deeply flawed, but simple to calculate and interpret, measure. This mistake of using a single simple but flawed metric to analyze accuracy pervades the logistics and supply chain management academic literature as well, as the plurality of reviewed articles chose mean absolute percent error (MAPE) to

evaluate forecasts. This metric suffers from inflation at low levels of demand, tends to drive systematic under-forecasting, is susceptible to outliers and is not estimable when demand is zero (Armstrong and Callopy 1992, Fildes 1992, Makridakis 1993, Hyndman 2006, Hyndman and Koehler 2006, Kolassa and Martin 2011). The primary concepts discussed in our literature sample were the tradeoffs of measuring bias and accuracy, comparability of measures and the insufficiency of individual traditional error metrics.

Though error metrics tend to command a greater portion of focus in both forecasting practice and research, and there is some indication that small nonzero bias in either direction can have positive effects on supply chain operations (Kolassa and Martin 2011, Sanders and Graman 2009), several papers indicate that reducing bias is relatively more important than reducing error to logistics and supply chain performance. Ebrahim-Khanjari et al. (2012) show that positive bias in shared forecasts deteriorate trust between wholesalers and retailers to a greater extent than error. Barman et al. (1990), Ritzman and King (1993), Flores et al. (1993) and Huang et al. (2011) show bias dominates error in its effect on cost reduction and delivery performance in production control systems, particularly with large lot sizes and small buffers, and that the positive effects of reducing error are due primarily to the reduction of the bias component of error. The same relative effect size is demonstrated in labor and inventory costs in a warehouse (Sanders and Ritzman 2004a, Sanders and Graman 2009, 2016), distribution network costs (Lee et al. 1993, Zhao and Xie 2002), and manufacturing supply chains (Chang and Yeh 2012). Bias measured cumulatively has been found to have a greater effect on logistics and supply chain costs for intermittent demand items (Willemain et al. 2004), and under some

control policies in a production control system (Sourirajan et al. 2008). Though Sourirajan et al. (2008) find that control policies that depend on cumulative information are more sensitive to bias, whereas rate based control policies are more sensitive to error outliers. Most of these works at least mention a tradeoff between bias and accuracy. That is, maximizing a measure of accuracy at best does not produce optimal bias, and can end up increasing negative effects of bias (Huang et al. 2011).

Barman et al. (1990), Syntetos and Boylan (2010) and Katsikopoulos and Syntetos (2016) present the tradeoff of bias and variance more explicitly as a reformulation of mean squared error (MSE), the only error measure that is tractably decomposed into bias, forecast variance, and random error variance components. This implies that there is a theoretical limit to how small MSE can be, as random error is by definition unexplainable. It follows that logistics and supply chain costs are minimized only by balancing bias and error terms (Katsikopoulos and Syntetos 2016).

While it may seem an attractive and simple way to benchmark performance, most metrics cannot effectively be compared across products, forecasters, forecasting methods or over time. This is particularly true of scale-bound metrics like mean error (ME), mean absolute error (MAE) (Moon 2015) and MSE (Chatfield 1992, Armstrong and Fildes 1995, Hyndman and Koehler 2006), but this applies to other error (and bias) metrics as well. Since the time series characteristics (not to mention other significant production and distribution considerations) differ between scenarios, we should not expect a common metric to hold the same meaning.

Three primary alternatives have been proposed to address limited comparability, and some eschew the use of accuracy measures altogether. The first alternative measures are nonparametric comparisons such as ordered or percent better measurements (Syntetos and Boylan 2005, Hyndman and Koehler 2006). However, such nonparametric comparisons can have low construct validity and deviate from the consensus of a number of other measures (Armstrong and Callopy 1992), making them potentially misleading in isolation. Accuracy implication metrics and loss functions are the second alternative that try to estimate the direct impact on the firm or supply chain (Lee et al 1987, 1993, Fildes 1992, Flores et al. 1993, Boylan and Syntetos 2006, Hyndman and Koehler 2006, Syntetos et al. 2010), with forecast accuracy metrics possibly not even included as an input. This may not be useful for diagnosing forecast misspecification, as accuracy metrics often have little relation to loss function or cost performance (Lee et al. 1987), or could have highly disproportionate effects on accuracy implication metrics with even small changes in accuracy (Syntetos et al. 2010). A third suggestion by Moon (2015) is to solely compare error rates over time, but even that admittedly assumes stationarity of demand drivers.

Comparability of metrics is even lower between different principals in the supply chain. Though many different may forecast demand, the level of aggregation, time horizon, and ultimate use of the forecast require separate metrics and result in different levels of both achievable and desirable accuracy. In a case study of a large industrial production firm, Kerkkänen et al. (2009) demonstrate that different metrics ought to be generated and utilized by different organizations generating forecasts. Sales, marketing,

finance, accounting, production, purchasing and logistics should all use metrics tailored to their function. This all implies that the "goodness" of any metric is less important than the metric being matched to the appropriate context. Hançerlioğulları et al. (2016) show that optimization of metrics at one level can negatively impact metrics at another in a study of financial data from 304 firms. They find negative bias, or what they call "sales surprise" a boon for investors in the short term by increasing inventory turnover, but that this negatively impacts production and long term customer goodwill. Conversely, they find increased forecast accuracy reduces turnover and short term shareholder value. This means that "goodness" of a metric is relative to the affected principal, and a firm, or a coordinated supply chain should balance the "goodness" of the most appropriate metrics between the multiple generators and beneficiaries of forecasts.

Finally, even metrics appropriately applied are found to be insufficient to completely summarize a demand forecast's quality. Knowing the level of error does not indicate the cost of such error, the persistence of error, whether positive or negative deviation affects the supply chain in different ways (Kolassa and Martin 2011), and whether errors affect different parties in a supply chain the same way. Several works present typologies of various forecast error metrics (Armstrong and Callopy 1992, Mathews et al. 1994, Hyndman and Koehler 2006, Kolassa 2016b), and many others describe potential defects and strengths of each measure. We loosely group our discussion by the types identified in Kolassa's (2016b) framework.

The first two types, absolute and squared errors, are both scale dependent, and so have low reliability (comparability between forecasts) (Hyndman and Koehler 2006).

Additionally, absolute error metrics such as MAE and Median Absolute Error (MdAE) rewards point forecasts for approaching the median, and so are sensitive to outliers (Kolassa 2016b). Squared terms such as the sum of squared errors (SSE), MSE, and root MSE (RMSE) rewards forecasts that approach the mean, are sensitive to outliers, exhibit non-normality and low construct validity (tend to disagree with the consensus of forecast measures) (Armstrong and Callopy 1992, Fildes 1992, Kolassa 2016b). Percent error measures like MAPE, median symmetric and weighted variants of MAPE, as well as mean percent error (MPE) each suffer in varying degrees from the limitations attributed to MAPE at the beginning of this theme. Relative error metrics, that use a benchmark forecast method (usually some variant of a naïve forecast) as a point of comparison share some inflation and definability shortcomings with percent errors (Makridakis 1993, Hyndman and Koehler 2006), scaled errors penalize over-forecasts more than under-forecasts (useful for intermittent data) but increase reliability, and loss functions tend to be highly context specific, complex (not easily interpreted) and non-generalizable (Armstrong and Fildes 1995, Kolassa 2016b).

Insufficiency of traditional measures has led to two basic propositions. First is to scrap traditional point forecasting methods in favor of predictive distributions, which then require alternate means of reporting forecast quality. The second is to account for the biases and informational deficiencies of single error metrics by including multiple metrics to balance them.

Adopting the first proposal, Kolassa (2016b) demonstrates by forecasting over 2000 products in drug and grocery retailers that accuracy metrics in common use are

inadequate, particularly for items with intermittent demand. Further, the use of a normal assumption when applying error metrics to determine safety stock ignores the asymmetry of error effects from over and under-forecasting. Kolassa suggests distributional forecasts rather than point forecasts to make up for the deficiencies he describes for each individual error measure in his typology. Point forecasts of all types provide too little information, particularly for very low and high demand values and skewed distributions. Zhao and Xie (2002) support this call for using predictive distributions as they found the type of forecast distribution, independent of either error and bias, significantly affected costs in a simulated distribution network. Limits to using predictive distributions instead of more traditional error metrics include the significant increase in data collection, required demand planner skill, reduced interpretability (Kolassa 2016b) and bias introduced from equal observation weighting in goodness of fit tests for distribution (Boylan and Syntetos 2006).

Mathews and Diamanopoulos (1994) and Wallstrom and Segerstedt (2010) adopt the second proposal. Mathews and Diamanopoulos find through principal components analysis (PCA) that 14 error measures reduced to four distinct components accounting for more than 85% of forecast variance. Ratio, volume, bias and fit-based measures (roughly equivalent to percent, scale dependent, bias and the coefficient of determination respectively) all account for different aspects of demand forecast variance. Also though PCA of forecast error measures for smooth (low CV, frequent), intermittent (low CV, infrequent), erratic (high CV, frequent), and lumpy items (high CV, infrequent), Wallstrom and Segerstedt suggest traditional error measures (MAE and MSE

specifically) insufficiently describe the difference of predicted and observed phenomena. Specifically, they find mathematically distinct and complementary explanatory power in MSE, MAE, sMAPE (symmetric MAPE), CFE (cumulative forecast error), PIS (periods in stock), and NOS (number of shortages). Morlidge (2014b) supports the approach of including multiple metrics in his examination of 11,000 demand forecasts of items with varying volumes and variance. He found that even though traditional measures of accuracy tend to be lower among high volume (and cost) items, they perform the same as low cost, low volume items on his proposed measure of RAE. This implies that traditional measures motivate a misalignment of resources, as high volume, high RAE items currently compose the largest possibility for cost improvements for demand forecasters. Reliance on a single or small number of measures leaves demand planners blind to potential areas for forecast improvement.

The tradeoffs between metrics means that the question "How good is good enough?" depends on the metric(s) in use and the type of forecast it is applied to. In general, positive and negative deviation will generate distinct cost (and revenue) effects that differ for each member of a supply chain. Extant literature suggests that error minimization can increase bias, which may increase overall costs. It is also suggested that regardless of what is considered "good enough" for a single product, that metric ought not be applied across all products, as differing circumstances and demand patterns make such comparisons inappropriate. "Good enough" should instead be based on more comparable nonparametric functions, accuracy implication metrics, loss functions, or solely be relative to past performance (assuming the demand pattern is more or less

stationary). Different metrics, and thereby different determinations of what is "good enough" should be applied based on the creators and consumers of each demand forecast. Marketers may be more interested in aggregate sales over a several month horizon, whereas operations may be more interested in production and distribution demands for a single product over several days. The negative impact of horizon and positive impact of aggregation on achievable accuracy in such differing applications are discussed in separate themes. Finally, research on the tradeoffs of metrics in logistics and supply chain management literature indicate that traditional error metrics are insufficient to communicate all of the relevant information regarding the difference in predictions from reality. Predictions ought to capture distributions, rather than point forecasts, and multiple measures are needed to convey all of the relevant information to decision makers, each with their own level of "good enough".

Future research on the tradeoffs of metrics must focus on quantifying the additional costs and potential benefits of following the suggestions of extant research. Shifting the focus of measurement from simple error to bias, or to a more holistic complement of measures for forecasts with higher information density will inevitably cost more in data collection, information sharing between supply chain partners, information technology (IT) investment, and recruitment and training for demand planners.

Level of Aggregation and Hierarchy

This theme can be viewed as a tradeoff of the *level of aggregation* at which the forecast is generated, and the *level of hierarchy* for which the forecast is generated.

Aggregation can be accomplished by product characteristic (Fliedner and Lawrence 1995, Moon 2015, Paul et al. 2015), location (Mentzer and Cox 1984, Zotteri et al. 2005, Williams and Waller 2011b), time (Zhao et al. 2002b, Rostami-Tabar et al. 2013, Jin et al. 2015), and even by customer profile (Chung et al. 2012, Breiter and Huchzermeier 2015), and consistently has been found to have a positive effect on the level of achievable accuracy. However, there exists a gap between theory and practice in aggregation methods. Practitioners tend to aggregate by predetermined hierarchies rather than covariance that would suggest effective grouping, which leads to mixed results in aggregation in practice (Syntetos et al. 2016). Aggregation typically also implies that forecasts are generated centrally (hierarchically), as upper echelons will have access to greater resources, more concentrated talent, and increased data quantity and quality. This also leads to greater degrees of accuracy in both demand and supply forecasts (Mentzer and Cox 1984, Davis et al. 2016), though this effect degrades for very short term forecasts and beyond the firm-level unit of analysis.

However, just because forecasts generated at a higher level of hierarchy or aggregation are likely more accurate, does not mean that they are more useful. In the end, the forecast must be tailored for the appropriate user, wherever they are in the hierarchy. Reconciliation of forecasts between functions, as well as between hierarchical levels remains a pressing issue for forecasters (Syntetos et al. 2016). A significant amount of logistics and supply chain management research aims to determine whether an automated means of reconciliation through a top-down strategy, in which aggregate forecasts are developed and later apportioned for use at lower levels of hierarchy,

provides better accuracy (and supply chain performance) than a bottom-up strategy, where disaggregated forecasts are later combined for use in higher levels of hierarchy. Zotteri et al. (2005, 2014) suggests that the choice between these demand forecasting strategies depends primarily on the relative import of sampling or specification error, and that minimum error is achieved through some mixture of these approaches.

Sampling error occurs when time series samples are short, noisy, nonstationary, contain errors, or include unrepresentative observations. Bottom-up approaches suffer when this type of error is common. Widiarta et al. (2006), Chen and Blue (2010), Hatzakis et al. (2010), Williams and Waller (2011b) and Rostami-Tabar et al. (2013) all find examples of this limitation when aggregating positively autocorrelated (within a demand pattern) or covarying (between demand patterns) demand signals. Williams and Waller (2011b) and Jin et al. (2015) demonstrate that though bottom-up approaches more often dominate top-down, limited disaggregated data can make bottom-up approaches less accurate than top-down approaches.

Specification error occurs when disaggregated units of analysis contain heterogeneous demand signals that are obfuscated by pooling demand, and thus top-down approaches suffer in the presence of this type. Fliedner and Lawrence (1995) demonstrate this in testing grouping mechanisms for diesel engine production spare parts forecasting. Flores and Wichern (2005) demonstrate that this confounding has a particularly acute effect on measures of bias. Caniato et al. (2005) show that a top-down clustering approach to account for structural, managerial and random irregularities can help save money on the (often) high costs of collecting diverse data, but that clustering

the random irregular data did not improve forecasts. This implies random effects wash each other out and represent misspecification. Moon (2015) presents an alternate example where relevant information is lost when grouping products that differ substantially in profit contribution. In this case accuracy may indeed be improved, but any sort of explanatory power for identifying causes of error in the relatively more important products is masked.

"How good is good enough?" depends on both sides of the tradeoff in this theme. Aggregation has been found to increase achievable accuracy, but forecasts must be tailored to the proper level of hierarchy for use. Tailoring the proper level of aggregation can be accomplished via top-down disaggregation or bottom-up aggregation, but the effectiveness of each approach depends on the relative import of sampling and specification error to a demand planner.

Future research on the effect of aggregation and hierarchy should attempt to differentiate the effects of different types and interactions of sampling and specification error on the effect of aggregation on demand forecast accuracy. For instance, if demand signals are both noisy and heterogeneous, is specification error or sampling error more significant. Also, we should expect the relative effect of these error types to differ based on the mechanism of aggregation. Future work should estimate how temporal, geographic, product and customer based forms of aggregation are affected differently by sampling and specification error.

Data Quality and Availability

Data quality and availability have been shown to be critical limiters of achievable demand forecast accuracy. The logistics and supply chain management literature on this theme cover both aspects of data input with regards to their effect on demand forecast accuracy.

Quality refers to errors in recording, undocumented substitutions, inventory record inaccuracy, data dependent on already included information, and erroneous or perhaps even dishonest reporting of data between supply chain principals. Quality has been found to suffer when policies for collecting data are nonstandard, poorly enforced, and involve judgement from personnel with low levels of familiarity or training. Such conditions have been found to be prevalent in humanitarian logistics operations (van der Laan et al. 2016), but varying degrees of these conditions have been found to be prevalent in demand management functions across all kinds of industries and situations. Sanders and Manrodt (2003) found that among 240 firms, forecasters were equally likely to pursue a qualitative forecasting strategy with subjective inputs, despite this approach achieving lower forecast accuracy regardless of industry, size or marketing strategy. They found the effect of data quality on accuracy was moderated by both forecasting expertise and degree of supply chain integration. This implies that the level of information sharing and coordination between members of a supply chain can also affect data (and therefore forecast) quality. Cachon and Lariviere (2001) demonstrate this analytically, and find that poor information quality not only negatively affects forecast accuracy, but that inaccurate shared information in a supply chain can erode trust and

eliminate the possible benefits from forecast information sharing. Clemen and Winkler (1985) find also that data quality (and its value to a forecast) deteriorates monotonically with the level of dependence the information has on previous data.

Availability refers to the lack of a capability to collect certain types of information, whether it be a technological, cost, access, or legal or regulatory limitation. Fildes and Petropoulos (2015) and Moon (2015) indicate that practitioners cite availability of or access to data as their most critical forecasting deficiency. This implies that demand forecasting processes, rather than technical capabilities, promise the greatest opportunities for improvement. As is the case with data quality, availability of data is lower in rapidly changing situations with fewer standard policies for data gathering and management (van der Laan et al. 2016). Kelle and Silver (1989) and de Brito and van der Laan (2009) find that reverse channels in closed loop supply chains often suffer from data availability, as data collection often depends on customers being incentivized to provide record rather than trained employees being compelled. Data collection in this case may also involve significant IT investment or the use of potentially costly tracking technology, which has mixed results on forecast accuracy (de Brito and van der Laan 2009). Holmström et al. (2006) find that in complex extended supply chains, different members of the supply chain may have different capabilities to generate and share accurate data, particularly at different stages of product, industry and firm maturity, as well as channel mix. Sanders and Manrodt (2003) note that differing levels of IT investment in firms also lead to lower levels of available data for qualitative forecasts at various levels of a supply chain.

For demand planners trying to answer "How good is good enough?", this means that there are limits to both achievable and desirable demand forecast accuracy as a result of these two dimensions of data inputs. Data can be of a lower quality due to unstable circumstances, which would not be possible to mitigate with a reasonable amount of investment. Availability may be lower due to the lack of technology to collect certain types of information. These would both decrease achievable forecast accuracy. On the other hand, some levels of quality and availability can be controlled, and end up as a management decision of what level of accuracy is desirable. Data quality that is driven by levels of demand planner training, enforced collection standards, IT sophistication, and quality control is strongly influenced by overall levels of investment. Availability can also be increased by incentivizing accurate information sharing with supply chain partners, or investments in technologies to increase data sources. It is up to the managers in a firm and across a supply chain to weight the costs and benefits of data quality and availability on demand forecast accuracy.

Beyond estimating how less accurate and more limited data affect forecast accuracy, the logistics and supply chain management literature has little that describes the effect of data quality and availability on overall business or supply chain performance. Future research in this area must attempt to quantify the effect of investments in quality or availability on more than just forecast accuracy (particularly single measures). This may include tying such investments into an integrated value model like the Economic Value Added (EVA) model presented in Lambert and Pohlen (2001), or even just counting it as a relevant cost in a decision model of tradeoffs.

Driver	*Implications*	*Proposed Future Research*
Forecastability	Low forecastable demand patterns have lower achievable accuracy, indicating "good enough" is also lower. May be better addressed with relative measures to target limited resources to only those demand forecast accuracy increases that promise the greatest return.	Examine interactions between multiple low forecastable situations, develop detection and avoidance mechanisms when forecastability is low.
Horizon	Where longer-range forecasts are required achievable accuracy is lowered, implying a lower "good enough". Where temporal aggregation is not possible to offset the lower levels of accuracy, alternate strategies such as postponement may be a better use of resources.	Compare the effect of forecast horizon in more varied contexts, and increase the estimation of nonlinear models.
Overfitting and Misspecification	Overfit models can seem more accurate than they end up being, particularly when estimating (and extrapolating) weak effects. Lower desired accuracy (by reducing training samples and increasing holdout samples) implies a lower level of "good enough". Resources are best spent training and recruiting forecasters with a sound understanding of modeling practices who are able to recognize overfitting.	Must focus on more effective detection of overfitting and of mapping the tradeoff of training and holdout sample sizes, particularly in cases of nonstationarity.
Tradeoffs of Metrics	"Good enough" depends on the metric in use, as each measure imposes different penalties depending on the type of deviation. Single metrics are inappropriate for comparison, and traditional error metrics insufficiently indicate the business impact of accuracy.	Increasing the number and complexity of measures to evaluate likely increases cost of demand planning, so benefits and costs should be quantified.
Level of Aggregation and Hierarchy	"Good enough" depends on the level of aggregation (which increases accuracy) and level of hierarchy at which the forecast is used. Tailoring aggregation to required level of hierarchy depends on the relative importance of sampling and specification error.	Should explore relative import of sampling and specification error when both are significant, and over multiple mechanisms of aggregation.
Data Quality and Availability	Data quality and availability limitations make "good enough" lower. Improving quality may be either impossible or cost prohibitive, and improving availability may be limited by technology. Resources can best be used to train the workforce on standard and accurate data collection to improve quality, and relationship management to increase information sharing to improve availability.	Should estimate the effect of data quality and availability on business performance measures such as EVA rather than (especially singular) measures of demand forecast accuracy.

Table 16: Summary of Technical Drivers of Demand Forecast Accuracy

Managerial Drivers of Forecast Accuracy Bounds

Managerial drivers are more under the control of managers and demand planners; exhibiting a greater degree of endogeneity than technical drivers. They represent differing degrees of capability or strategic emphasis on forecasting accuracy. They include policy considerations, various interest tradeoffs, and inter-firm dynamics that primarily affect the level of desired forecast accuracy. Literature findings, implications for "good enough", and proposed future research is discussed below for each managerial driver of forecast accuracy bounds, summarized in Table 17.

Error Amplification

Error has long been known to amplify as a demand signal moves back up a supply chain, in an effect known as the bullwhip effect. This means that at higher echelons, achievable forecast accuracy is typically lower. While most research on the bullwhip effect involves integration and information sharing efforts to mitigate this amplification, we find three groupings within the theme of error amplification which either enhance or diminish the importance of error between echelons.

First, there are situations where the effects of error amplification are expected to be particularly acute. This implies that requirements for "good enough" are perhaps less achievable, and that alternate means of mitigating extreme bullwhip may be more effective. The case of no information sharing or inaccurate information sharing is discussed in separate themes. Lorentz et al. (2007) find that underdeveloped distribution networks with unreliable infrastructure, carriers, and suppliers exacerbate the bullwhip effect. The mitigating effect of information sharing is enhanced in this context, though

through informal channels and individual relationships (increasing risk of corruption and extortion). van der Laan et al. (2016) find similar results in humanitarian logistics operations due to obsolescence of shelf-life items, highly fractious supply chains with unintegrated agencies, short working histories and unstable supply. This also results in potential underreporting and hoarding due to scarcity.

Second, we found work that examined effects other than error amplification apparent in bullwhip. This implies that reducing error or error amplification will not fully eliminate the damaging effects of bullwhip, relaxing requirements for "good enough". Sanders and Graman (2016) find significant bias magnification in addition to error in a retail distribution simulation corroborated with survey responses. Hosoda and Disney (2006) find inventory variance increases alongside demand variance in a cost model of a manufacturing supply chain, and find that error reduction does not reduce supply chain costs when inventory variance inflation exceeds demand variance inflation. Ma et al. (2013a) similarly find significant costs from inventory, not demand or production, variance inflation, and discuss the differential effects between supply chain members. Bullwhip costs from inventory oscillation costs more for downstream members, while upstream members bear more costs from production oscillation. This indicates that upstream supply chain members benefit unequally from efforts to reduce demand variance inflation.

Third, several research papers examined the relative role of demand forecast error reduction in combatting the ill effects of bullwhip. This indicates that "good enough" may not be as important as other factors, and may therefore be relaxed. In testing the five

drivers of bullwhip proposed in Lee et al. (1997), Agrawal et al. (2009) find that information sharing (and therefore forecast accuracy) is dominated by lead time in bullwhip reduction, and that some level of bullwhip is unavoidable. Doganis et al. (2008) explore the bullwhip effect in the Greek dairy industry, and find that forecast error reduction outpaces cost reduction. This implies diminishing returns on forecast accuracy investments for upstream partners. Williams and Waller (2010) found that in grocery store replenishment, if non-turn volume or information distortion was high, distributors benefitted more from using order (rather than POS) data. This implies forecast accuracy has a natural limit in instances where internal dynamics that distort information are high. Lackes et al. (2016) find analytically that contract penalties for downward order revisions can replace upstream forecast accuracy via POS information sharing. Though in price or buyback penalty contracts, suppliers are found always to benefit from increases in forecast accuracy (Amornpetchkul et al. 2015).

Currently, the logistics and supply chain management literature indicate that while error is important to the costs borne from bullwhip amplification, there are other significant effects and potential solutions to the costs of bullwhip. Even in contexts that are prone to error inflation, considerations like bias and inventory inflation must be weighed with error inflation, and may even deserve higher priority depending on industry context and level of hierarchy. Reducing the costs of bullwhip can also be achieved through means other than demand forecast error reduction, such as through lead time reduction, contract incentives, or realignment of internal drivers of information distortion. "How good is good enough?" is unfortunately complicated by all of these issues, and

practitioners must assess their susceptibility to bullwhip, the relative import of the error amplification component of bullwhip costs, and feasibility of alternate means (other than error reduction) to combat these costs.

Future research should aim to assist this effort. While efforts in the logistics and supply chain management literature to date address identification of at most small numbers of bullwhip cost drivers or mitigators, they must shift to more comprehensive diagnoses and normative responses.

Cost Tradeoffs

While many of the themes previously discussed involve cost tradeoffs, the papers in this theme address costs explicitly, and across a more diverse array of potential sources in the supply chain. This includes works that discuss costs associated with increasing demand forecast accuracy, relative direct and indirect costs resulting from forecast error, offsetting costs affected by error, compensatory costs which may dominate the effect of error, major production policy considerations of error levels, non-pecuniary error costs that have been considered, and the unequal distribution of costs and benefits between supply chain members.

In any discussion of "How good is good enough?", there must be an accounting of what it would take to improve on the current state of demand forecast accuracy. Improvements rarely come free, and tend to be nonlinear and discrete with each type and level of forecasting investment. Despite overwhelming evidence that such investments as IT integration and systems to generate statistical forecasts improve accuracy, nearly half of firms use simple spreadsheets to store demand data and generate forecasts (Canitz

2016), and firms are equally likely to use simple qualitative forecasting methods as more accuracy quantitative methods (Sanders and Manrodt 2003). Numerous costs related to improving forecast accuracy act as barriers to pursuing more accuracy forecasts. Recruiting forecasters with technical talent or providing better training forecasters improves accuracy (Mentzer et al. 1984, Chung et al. 2012, Canitz 2016, Doering and Suresh 2016, van der Laan et al. 2016), and such training is readily available through professional organizations as APICS, CSCMP, IBF and IIF. Accuracy can be improved through investments to collect and store additional data such as customer profile information (Chung et al. 2012), data collection technology, infrastructure or incentives (Kelle and Silver 1989), exogenous environmental factors for integration into forecasts (Trapero et al. 2012), product performance and failure data (Tibben-lembke and Amato 2001), and data storage capacity to store it all (Sanders and Manrodt 2003). Accuracy investments may also include more sophisticated forecast support systems for generating, tracking and managing forecasts (Trapero et al. 2012, Canitz 2016, Doering and Suresh 2016). In a supply chain, investments for sharing information are necessary for improving forecasts upstream (Kelle and Silver 1989, Chung et al. 2012, Trapero et al. 2012, Babai et al. 2013, Canitz 2016, Lackes et al. 2016, Huang et al. 2017), and involve potentially unbalanced costs between supply chain members. Each of these costs may include some discrete up-front investment as well as maintenance, training and continuation investments that differ by industry, product and supply chain relationship, making it difficult for a firm to determine a marginal cost for accuracy improvement. Additionally, such investments do not necessarily guarantee any measurable level of

improvement, merely greater forecasting capability. Actual accuracy improvement depends on many (and in some cases all) of the themes we identify in this paper, though increased capabilities have been shown to improve supply chain cost and service even when accuracy is not improved (Doering and Suresh 2016).

To motivate investment in increased accuracy, what costs are relevant to consider? We reviewed logistics and supply chain management literature that indicates direct and indirect costs from demand forecast error, as well as costs that offset other relevant costs but are also affected by levels of error. Forecast error has been found to drive higher unfinished, buffer and finished inventory carrying costs (Schmidt 1984, Wemmerlöv 1984, 1989, Sridharan and LaForge 1989, Ritzman and King 1993, Zhao and Lee 1993, Tibben-lembke and Amato 2001, Kahn 2003, Sethi et al. 2007, Kerkkänen et al. 2009, LeBlanc et al. 2009, Persona et al. 2011, Chang and Yeh 2012, Babai et al. 2013, Ma et al. 2013a, van der Laan et al. 2016), higher production and storage capacity costs (Schmidt 1984, Kerkkänen et al. 2009), higher switching, freezing, lot sizing or other production instability costs (Schmidt 1984, Sridharan and LaForge 1989, Zhao and Lee 1993, Lin et al. 1994, Kahn 2003, Sethi et al. 2007, Yelland 2010), higher lost primary and accessory sales, ordering, transportation, trans-shipping and expediting costs due to shortages (Wemmerlöv 1989, Tibben-lembke and Amato 2001, Kahn 2003, Sethi et al. 2007, LeBlanc et al. 2009, Persona et al. 2011, Chang and Yeh 2012, Canitz 2016), higher trans-shipping and lower margins from overages (Kahn 2003, Chang and Yeh 2012), higher design and production costs from flexibility and postponement response strategies (Wemmerlöv 1984, 1989, LeBlanc et al. 2009, Yelland 2010), higher

warehousing costs (Sanders and Ritzman 2004a, van der Laan et al. 2016), higher spoilage or obsolescence costs (Kahn 2003, Chang and Yeh 2012, van der Laan et al. 2016), and higher information sharing costs (Yelland 2010).

Some indirect costs come as a result of higher forecast error reducing service level, reducing customer satisfaction, increasing lead time, reducing value of information sharing, and reducing growth and shareholder value (Wemmerlöv 1984, Kahn 2003, Canitz 2016). Though these costs are harder to measure and are certainly influenced by other factors, they are the most critical to securing top-level support for any major investments in demand forecast accuracy. Without demonstrating the effect of forecasting accuracy on shareholder value, such improvements will be up to operational level leaders with fewer resources available.

The challenge of estimating the effect of a change in accuracy is the dynamic responses from various offsetting costs. Both hard and soft constraints may drive higher requirements for offsetting compensatory costs. Investments to increase accuracy can reduce reliance on such compensatory measures, but likely in an asymmetrical manner. For instance, increased service level requirements may be achievable through either flexible production or increased inventory (Wemmerlöv 1984). Increased forecast accuracy can also help achieve that service level, while simultaneously reducing the required investment in flexibility or inventory. The difference is that flexibility investments may be more capital intensive, and therefore have large discontinuities, whereas inventory reductions would more likely be incremental. This implies that when weighing the cost tradeoffs of forecast accuracy improvements, managers must consider

multiple real options for alternative compensatory investments. Realized cost savings, however, will only be from the option exercised. Huang et al. (2011) demonstrate this effect as they show lead time reduction replacing the need for accuracy to offset flexible production or increased inventory costs. Sridharan and LaForge (1989), Tibben-lembke and Amato (2001) and Persona et al. (2011) show this in offsetting costs of inventory and lost sales, representing opposing signs of forecast bias, that are simultaneously reduced with reductions in error. The tradeoff of buffer inventory costs and production switching costs is shown to be moderated by forecast accuracy (Zhao et al. 2001). Yelland (2010) demonstrates this in the tradeoffs of information sharing costs and flexible manufacturing costs, as reliance on both are reduced by forecast accuracy.

Some studies have found these offsetting costs to be more effective than forecast accuracy improvement at reducing overall costs. Sridharan and LaForge (1989) found inventory savings to be greater from investments to reduce setup time than similar investments to improve forecast accuracy. Ritzman and King (1993), Clark (1998) and Venkataraman and D'Itri (2001) found costs associated with matching lot sizing and inventory policies was more effective than forecast accuracy investments in reducing inventory and production costs, though this is not supported in all cases (Fildes and Kingsman 2011). Sanders and Ritzman (2004a) indicate warehouse workplace flexibility can offset accuracy, though this is not as effective against bias. Finally, some cost savings typically associated with error reductions are truly a result of reductions in bias. As Chang and Yeh (2012) note, over-forecasting costs such as excess inventory carrying cost, trans-shipping cost, obsolescence, reduced margin and labor; and under-forecasting

costs like expediting, higher per-unit production costs, lost sales, lost accessory sales, reduced customer satisfaction, and delayed delivery are most often conflated as simply error costs.

Beyond moderating a tradeoff of two or a few offsetting costs, achievable accuracy levels may even dictate production control methods, involving differing costs associated with aggregate production, labor, material, lot size, holding, shortage and switching costs. Such drastic changes in cost basis have been observed in multiple contexts, often involving estimation of policy tradeoff curves. Barman et al. (1990) show that linear decision rules for production shifts, optimal under low or no forecast error, are dominated by a proposed production switch heuristic when error is nontrivial. Johnson and Anderson (2000) generate tradeoff curves for transition to a postponement production strategy in the HP printer supply chain, dependent on forecast accuracy. Jeong (2011) estimates a supply chain decoupling point, or lateral position in a supply chain where production shifts from make to stock to make to order. The further back in the supply chain the decoupling point is, the greater reliance a supply chain has on forecast accuracy. High costs to accuracy, low achievable accuracy, high inventory costs, and greater demand for customization drives this decoupling point forward and thus reduces the reliance on highly accurate forecasts. Wemmerlöv (1989) demonstrate that high achievable accuracy is a prerequisite to just-in-time inventory minimizing production strategies.

Tradeoffs do not necessarily need to be of a pecuniary nature, as Niakan and Rahimi (2015) demonstrate in a healthcare supply distributor and Ji et al. (2014) in a

manufacturing system that public externalities in the form of greenhouse gas emissions, energy, and toxic chemical consumption can factor into a firm's decision to invest in forecast accuracy. However, these are increasingly becoming real costs through environmental regulation. van der Laan et al. (2016) demonstrate in humanitarian logistics operations that a particularly difficult proposition is determining the value of the potential to save a life by increasing forecast investment. Wacker and Sprague (1998) show that different cultural (as measured by Hofstedte's cultural dimensions) values in seven countries affect the required level of forecast accuracy and investment in technology. For instance, greater power distance and individualistic cultures are more likely to invest in demand forecasting technology and depend on more accurate statistical forecasting methods, masculine and individualist cultures are more likely to depend on qualitative factors and manual adjustments, and masculine cultures are more likely to involve senior leadership in forecast development. These non-pecuniary biases, altruistic impulses, cultural and societal pressures, and risk orientations (discussed separately) must be considered with cost tradeoffs of forecast accuracy, even if they are not directly implemented in a model of costs.

In a supply chain, the complex interdependent cost tradeoffs with forecast accuracy are also unequally shared between different members of a supply chain. This often introduces additional inducement, contracting, or information sharing costs to balance this asymmetry, moderated by the balance of relational power and level of information asymmetry. Hosoda and Disney (2009) and Chen and Xiao (2012) both show that increasing forecast accuracy to minimize local costs can actually increase

overall supply chain costs, as additional costs from moral hazard manifest from information asymmetry. Zhao et al. (2002b) and Sethi et al. (2007) both note that downstream members of the supply chain typically bear higher costs from forecast accuracy investments, and do not share equally in benefits of forecast accuracy. Upstream members typically benefit from increased accuracy through reduced production and distribution variance, whereas downstream partners benefit primarily from inventory reductions (Hosoda and Disney 2009, Lackes et al. 2016). Taylor and Xiao (2010) find this effect to be convex, and that manufacturers only enjoy cost benefits from accuracy improvements when downstream demand planners already have acceptable forecast accuracy. Miyaoka and Hausman (2008) show that cost benefits enjoyed by upstream supply chain members from increased downstream demand forecast accuracy are only realized if they set the wholesale price (either by enjoying greater market power or as the Stackleberg leader). Downstream members are motivated to over-order (Chen and Xiao 2012, Lackes et al. 2016), particularly in cases where buybacks or cancellations are permitted. In such situations, upstream supply chain members can bear the additional cost of incentives, contract costs to implement buyback or cancellation penalties (Zhao et al. 2002b, Sethi et al. 2007), or information sharing costs (Huang et al. 2017) in order to motivate forecast accuracy investments by downstream members in order to avoid excess production variance costs. Inducement costs can increase with information asymmetry between supply chain members, the relative power of principals, and the risk orientation of each (Lackes et al. 2016). Chang and Yeh (2012) propose a method for objectively controlling cost sharing in demand planning collaboration in a steel manufacturing supply

chain through a prior agreement on exception thresholds defined by Taguchi loss function for the supply chain, thus helping to minimize additional costs from moral hazard.

For implications on "How good is good enough?", cost tradeoffs are the most difficult to accurately (and completely) identify, but likely the most important in determining a final level of desired accuracy. Improving forecast accuracy can involve significant investments in forecaster skill, as well as IT tools for gathering data, generating and tracking accurate forecasts, and sharing relevant information with supply chain partners. Though even with such investments there is no guarantee of accuracy improvement, there are significant potential cost reductions from pursuing these capabilities. Managers must determine which costs are relevant to their situation, accurately measure them, and determine the potential improvements from incremental changes in forecast accuracy. Cost savings from increased accuracy can be direct, such as through reduced inventory or production variance costs, or indirect, by reducing those costs associated with lost shareholder value from lower customer satisfaction. Multiple offsetting potential costs must be weighed against forecasting investments, and in some cases alternate compensatory investments may be preferred to investments to improve accuracy. The achievable level of accuracy may also dictate the policies and types of investments necessary to meet demand at the lowest cost. Improvements in demand forecast accuracy can reduce non-pecuniary costs as well, as social, cultural and environmental considerations increase in importance to a firm. Finally, managers must take into consideration that the costs and benefits of forecast accuracy improvements are borne unequally between supply chain members. To maintain a healthy supply chain and

maximize the benefit for all, some supply chain members must bear additional costs to share the burden of accuracy improvement costs.

As with other themes, the considerations for cost tradeoffs merely identify the relevant costs and measured effects currently present in the logistics and supply chain management literature. To act on these findings, a firm needs to assess their own context and relevant costs. Estimating the cost tradeoffs of nonlinear and discontinuous accuracy investments with separate nonlinear and potentially discontinuous functions for various offsetting costs (or real options) with any degree of accuracy is a complex and difficult endeavor. However, armed with the knowledge from the other themes discussed in this work, managers can make an informed decision of desirable accuracy, given contextual factors that drive achievable accuracy. One helpful suggestion is to focus what is likely to be a significant effort on those products that promise the most return from forecasting accuracy investment. Morlidge (2014b) suggests, in their examination of forecasts from 11,000 products, that as little as 6% of products make up nearly two thirds of the potential improvement in forecasts due to forecaster skill (and 40% of forecasts would be made better by simply implementing a naïve forecast).

Future research should focus on aiding practitioners navigate the complex interdependencies of estimating cost tradeoffs. Extant research seems somewhat myopic in examining relatively few local cost tradeoffs. Researchers must strive to capture a more holistic set of interdependent costs and opportunities. Canitz (2016) suggests that forecasting investments will remain low until firms can demonstrate value exceeding additional costs through such tools as the Du Pont Financial performance model, or we

would propose EVA, as suggested in Lambert and Pohlen (2001). Research must target research to aid this effort.

Supply Chain Integration

We found in our search of logistics and supply chain management literature that the initially separate search topics of "information sharing", "collaboration" and "integration" were largely conflated, and so were not effectively separable. Therefore, the emergent theme we call supply chain integration exists as a continuum spanning simple information sharing through complete integration of forecasting efforts in a supply chain. We separate the reviewed literature into work on the simple information sharing side of the spectrum from work that involves more direct collaboration and in some cases integrated demand planning.

First, we discuss the effects of various types of information sharing that have been studied in the logistics and supply chain management literature. Ramanathan (2013) describes multiple characteristics of information that can be shared between (and within) firms for the purpose of forecast accuracy improvement, including factors of importance, relevant forms of information shared between supply chain entities, and the implications of external sources of information. We use this typology to examine the literature on simple information sharing.

Factors of information importance include cost, which can be cost of analyzing, sharing, gathering, and how these risks or costs (and resultant benefits) are shared between supply chain members; usability, or the way it can or will be used; reliability, or accuracy of input data, trustworthiness and persistence of the source; actionability, or

how readily it can be converted to immediate utility; and finally capability, or the capacity of the information acquirer to transform the information to a valuable outcome. Eksoz et al. (2014) find these factors to be important to both departmental (internal collaboration) and group (external collaboration) information sharing.

As discussed in the previous theme of cost tradeoffs, costs and risks for forecast accuracy improvements tend to fall more heavily on those downstream organizations with direct access to demand. Cost savings also tend to be greater upstream in the supply chain for improved accuracy. This makes various types of demand information or demand forecasts valuable to upstream partners, who are then willing to share some burden in order to gain access to that information in order to optimize supply chain benefits. Ülkü et al. (2007), Zhu et al. (2011), Kurtuluş et al. (2012) and Bian et al. (2016) generalize this to show that increased cost of forecast accuracy increases the value of information to the supply chain, but disproportionately against whomever makes the investment and to whom is primarily responsible for inventory costs. Yao et al. (2005) provide an example of this more general view where the upstream manufacturer has greater forecasting capability than their customer. They find that for manufacturers employing a multichannel marketing approach with a direct channel, information sharing from a customer-competitor only holds value if the customer's forecasting capability exceeds the manufacturers'. In addition to the cost of accurate information, the level of information asymmetry between supply chain principals increases the value of information sharing. Kung and Chen (2014) find that upstream members of the supply chain benefit from sharing forecast information when asymmetry is high, but only if all

independent downstream members are synchronized in both their accuracy and sharing. Hartzel and Wood (2017) find support for the asymmetry argument when they observe greater value from POS information sharing when order frequency is low, and shared information is stationary and non-intermittent. This implies that upstream signals without sharing have low fidelity as a function of distortion, and shared POS data has high forecastability. Bian et al. (2016) also shows that in industries with high competitive intensity, competing supply chains also benefit from observing the actions of information sharing supply chains.

Cachon and Lariviere (2001), Terwiesch et al. (2005) and Guo et al. (2006) indicate that a lack of trust of information from downstream partners may hinder information sharing, particularly if information is volatile or found to have (particularly positive) bias. They find that upstream partners who normally benefit unequally from information sharing are unwilling to pay for incentives to share information if trust low, and instead institute penalties (delayed delivery, reduced capacity dedication) for untrustworthy downstream partners. Paradoxically, this increases volatility and positive bias in downstream forecasts, and further reduces the demand for accurate forecast information sharing.

Though the majority of logistics and supply chain management research on information sharing has to do with either point of sale demand data or demand forecasts, relevant forms can include sales data; order data; seasonal, discount, promotional and historical sales; relevant regulations and policies; and local forecast information. Sharing these types of information has been shown to simultaneously reduce the negative effects

of demand forecast error and increase the achievable forecast accuracy at higher echelons, as discussed above in the error amplification (Williams and Waller 2010) and aggregation (Williams and Waller 2011b) themes.

Sharing POS data has been found to lower upstream supply chain labor and inventory costs through improved replenishment forecast accuracy (Trapero et al. 2012, Babai et al. 2013) and reduced bias (Sanders and Graman 2016), and dampen the inventory costs from positive bias introduced by increased product variety (Wan and Sanders 2017). Though as Cui et al. (2015) note, POS information sharing alone may not improve upstream forecasts unless alternate information such as replenishment policy is also shared. Similarly, demand forecast sharing has been shown to reduce inventory and production costs (Zhao et al. 2002a, Ali et al. 2012), reduce demand variance amplification (Ma et al. 2013b), and can serve as a substitute for downstream demand forecast accuracy (Yao et al. 2005), each moderated by the degree of demand forecast accuracy.

Alternate forms of information to share include engineering and failure data (Tibben-lembke and Amato 2001), additional information on demand or forecasts such as the distribution (not simply the level) of uncertainty Terwiesch et al. (2005), relevant replenishment and stock policies (Williams and Waller 2011a, Cui et al. 2015), inventory and return data both between firms and between channels (Coronado Mondragon et al. 2011) that can serve as replacements for forecast accuracy. These alternate data forms can improve upstream performance even when accurate demand or forecast information is not shared.

External sources may include competitor and market data. Most of this would be gathered from a third-party marketing firm or consultant, and can be expensive. Achievable and desirable forecast accuracy improvements from external data such as this would depend on its availability, quality and cost.

Beyond simple information sharing, there are numerous examples in the literature of the effects of collaboration or full integration on forecast accuracy. McCarthy and Golicic (2002) and Yao et al. (2013) identify some of the major difficulties in collaborative planning, forecasting, and replenishment (CPFR) implementation. In the long-run it promises forecast accuracy improvement and lower costs, but implementation costs include software and technology, investments in coordination and information exchange, time and personnel for set up and maintained coordination operations, difficulty in scalability from pilot usage across suppliers and products, and perhaps most difficult synchronous change efforts in multiple firms. Improvements in forecast accuracy do not always align with inventory cost savings as the collaborative efforts are harmonized, and this also depends on product life cycle stage. There is also risk, as these costs do not guarantee successful implementation, and may disrupt and hinder (particularly replenishment) operations initially. Moon (2015) notes that the effectiveness of collaborative forecasting depends on the internal source of data, and both the generator and consumer of a forecast. If all three of these groups share interests and integrate effort, then collaborative forecasting can provide value. Nagashima et al. (2015) demonstrate that collaboration intensity improves forecast accuracy, but only among products that have low market competition (are inimitable). In one form of

integration, vendor managed inventory, Kannan et al. (2013) and Kurtuluş et al. (2012) show that such an arrangement shifts the cost burden for order processing costs, transportation costs, lot quantity costs, and in some cases inventory carrying costs upstream to a supplier. This significantly affects the desirable level of forecast accuracy based on the previously discussed sets of cost tradeoffs.

For the manager attempting to determine "How good is good enough?", various degrees of supply chain integration typically entail greater achievable accuracy at higher echelons of the supply chain, and lower required accuracy at lower echelons. This effect depends on several factors of information importance, such as the cost of the data, level of information asymmetry, the trustworthiness of the shared data, and for shared forecasts the accuracy of the downstream forecast. For organizations that are further along the spectrum and collaborate or are truly integrated in their demand planning, "good enough" can change substantially based on how the collaboration shifts cost burdens and on the effectiveness and maturity of the collaborative relationship.

As most work on information sharing focuses on sharing POS data and demand forecasts, it appears there is an opportunity to develop our understanding of variable effects of sharing different types of data. To date, little has been done to differentiate the effect by information type. Future research should also focus on how different collaboration types shift the complex set of cost tradeoffs that dictate "good enough".

Supply Chain Flexibility

Flexible strategies such as postponement or standardization of parts (as in a super bill of materials) can serve as a substitute for forecast accuracy, and in some cases

improve forecast accuracy. The value of postponement increases with increased forecast error (Johnson and Anderson 2000) and degree of product proliferation (Lee 1996), while greater degrees of postponement increasingly diminish the effect of forecast accuracy (Wemmerlöv 1984). Postponement and standardization of parts have been shown in improve forecast accuracy when the firm desires greater product variety (Persona et al. 2007, Paul et al. 2015), but the success of such strategies depends on the maturity of an industry, market, technology or product (Terwiesh et al. 2005), as well as the quality of management (Khouja and Kumar 2002). Revolutionary products likely do not have the market penetration to offer variants that permit postponement or standardization. By either increasing accuracy or replacing the need for accuracy, flexible strategies shift the relative cost tradeoffs with lower levels of forecast improvement investments (Khouja and Kumar 2002, Graman and Sanders 2009) and both unfinished and finished goods inventory (Persona et al. 2007, Graman and Sanders 2009), while increasing costs for sourcing (Paul et al. 2015), engineering (Persona et al. 2007), redesign (Lee 1996), postponement variety and volume capacity (Khouja 1998, Graman and Sanders 2009) and manufacturing of common parts (Khouja and Kumar 2002).

Implementing flexible strategies can mean "good enough" is lower, when used to replace relatively expensive improvements to forecast accuracy, or can alternatively mean "good enough" is higher due to increased achievable accuracy. This will depend on the perceived necessity of product variety.

Future research on flexible strategies in relation to forecast accuracy must address many of the same issues identified for future research on cost tradeoffs. To date,

considerations of such policies are myopically focused on local costs. The potential impact on value generation when shifting to flexible strategies means that research ought to include more cross-functional perspectives. Operations and logistics costs are currently examined in the research, but marketing and research and design considerations are either ignored or remain untested. As with other themes, the scope of research ought to span the supply chain rather than the firm, as accuracy implications of flexible strategies are likely unequal between different supply chain principals.

Manual Adjustments

Though generally statistical forecasts are found to have greater accuracy than subjective or qualitatively generated ones, subjective manual adjustments are found to improve upon statistical forecasts. The effect of manual forecast adjustment is even found to have an enhanced effect on statistical forecast improvement when incorporated as automated adjustment heuristics (Hur et al. 2004, Fildes et al. 2009). Such adjustments have limitations, however, as they can add considerable cost and time when there are many forecasts to generate (Sanders and Ritzman 1995). Manual adjustments can replace error with more costly bias (Fildes et al. 2009, Wan and Sanders 2017), especially when products are newly introduced or in the decline phase of their lifecycle (Petersen 2003), and lose effectiveness or can even increase error when adjustments are small (Fildes et al. 2009) or frequent (Wacker and Sprague 1998). The effectiveness of adjustments depends on degree of adjuster expertise, but also can differ by biases from personal motivation (Eroglu and Croxton 2010), organizational motivation, information access, level of procedural control (Oliva and Watson 2009), culture (Wacker and

Sprague 1998) and gender (Eroglu and Knemeyer 2010). Effectiveness may not be measured as reductions in forecast error or bias, as Ebert and Lee (1995) show manual adjustments can reduce production operation costs while decreasing accuracy. Effectiveness can also depend on the type of information introduced by an adjuster. Marmier and Cheikhrouhou (2010) note that time sensitivity, immediacy, impact and persistence of information all can affect adjustment quality.

In determining "good enough", this may mean accounting for the cost or availability of incorporating manual adjustments. If error is high, manual adjustments are more likely to improve accuracy. If there are numerous products to forecast with high frequency, adjustments can be expensive and may require contextual knowledge that might not be available. Though automated heuristic implementations of manual adjustments could increase speed or reduce costs, Fildes et al. (2009) note that these features are not yet common in forecast support systems, and may require significant technology investments. Managers must assess their situation to determine if they have a positive potential for forecast accuracy improvement from incorporating manual adjustments, and weight this against the realized costs from implementing these adjustments.

Research should support this effort to a greater extent. Manual adjustments ought to be examined on a contextual basis. The effectiveness of adjustments has been shown to depend on the size and frequency of an adjustment, and on numerous characteristics of the adjustor, but what about forecasts generated for different purposes, at differing levels

of hierarchy, and over different horizons? Academic research will provide the most benefit for practitioners if it can normatively prescribe action based on circumstance.

Risk

Risk occupies a separate theme from cost tradeoffs because most often cost tradeoffs imply perfect information and rational action. Situations of actual risk involve a significant lack or loss of information. Perceived risk drives irrational action based on risk preference (for both seekers and avoiders). For this reason, the investigation of the limits of forecast accuracy based on actual and perceived risks from forecast error occupies a significant stream of logistics and supply chain management literature.

Sanders and Manrodt (2003) identify environments likely to be sensitive and insensitive to demand forecast error that they call high and low uncertainty environments. High uncertainty environments include frequent product changes, short product lifecycles, market substitutes, global competition, low market power, and potential for obsolescence, and are likely to not only have higher degrees of forecast error but to also experience higher costs as a result. Low uncertainty environments include those with monopoly protections such as patents, long product life cycles, and unique products with high barriers to entry, and are unlikely to incur as great of costs as a result of forecast error.

Examples of high uncertainty environments include food supply chains (Eksoz et al. 2014), susceptible to spoilage and waste, but also risks to public health and from regulatory intervention. Thiel (2014) describes the massive shifts in both demand and supply when a health risk prompts a recall. In the case of a recall, hedging sources or a

flexible strategy of production dominates forecast accuracy, which may not be possible in unpredictable public risk perception. Rapid innovation and heavy regulation in an industry, as is experienced in the pharmaceutical industry, can also make rapid and dramatic changes to demand conditions commonplace (Kiely 2004). Lorentz et al. (2007) provides an example of poor infrastructure inducing variance in supply and production, making stable and efficient operations impossible and exacerbating the effects of forecast error. Wickramatillake et al. (2007) show that in the execution of major projects like infrastructure construction, the interdependencies of different suppliers, service providers, regulators and principals make such endeavors particularly susceptible to forecast errors.

Responses to high uncertainty typically entail some hedging or robustness strategy, as forecast investments likely have poor returns. Supply chains can be made more robust to high uncertainty through diversification of supply (Cachon and Lariviere 2001), excess inventory or production capacity (Georgiadis and Vlachos 2006, Kerkkänen et al. 2009), through investments in increased communication (Terwiesch et al. 2005), or through agility investments like dynamic assortment planning (Rajesh and Ravi 2015). As high uncertainty environments tend to have extremely low forecastability, in both demand levels and for relevant costs, there is a much greater potential for risk preference of decision makers to drive significant over and under investment in forecast accuracy and other hedging strategies.

The negative effect from perceived risk differing from actual risk can manifest in both risk seekers and risk avoiders (neutrality implies cost optimal decisions). Managers

who are risk seeking may choose not to hedge, instead seeking short term cost savings, but in the end face major shortage costs. Risk averse managers instead overspend on a risk management over time so that the large, but rarer, shortages can be avoided. These actions may actually cause greater uncertainty for other members of the supply chain. Guo et al. (2006), Chen and Xiao (2012) and Lackes et al. (2016) demonstrate that a more risk averse downstream partner is more likely to over-order (inducing bullwhip), and will likely charge a higher premium to share POS or forecast information, more valuable to their suppliers. Shin and Tunca (2010) find that firms can overinvest in demand forecast error reduction when perceived market competition is high, or if barriers to demand forecast improvements are low (implying competitors are likely to make such investments). Li et al. (2015) show that risk aversion leads to overinvestment in lead-time reduction with increasing forecast error, which can be somewhat mitigated with revenue sharing contracts between supply chain echelons. Risk aversion can be considered a significant, if difficult to measure, problem among logistics and supply chain managers as Smith and Mentzer (2010) find strong belief among managers that improved forecasts will improve decision making and logistics performance. While generally true, this belief can lead to risk averse overinvestment when specific conditions do not call for increased accuracy.

Correcting the gap between perceived and actual risk can be accomplished through increased communication, and strengthening of relationships between supply chain members. Ebrahim-Khanjari et al. (2012) and Gönül and Goodwin (2012) find that perceived competence, benevolence and integrity between retailers and suppliers fosters

trust and reduces the risk from forecast inaccuracy. Similarly, Bian et al. (2016) find trust to reduce the inequality of value from information sharing. That is, regardless of forecast accuracy, benevolent forecasters are trusted and selfish forecasters are not trusted (though error and bias affect future trust).

For the manager determining "How good is good enough?", the key task from this theme is determining the levels of actual and perceived risks they are exposed to. Higher levels of uncertainty in their environment mean achievable accuracy is lower. What is more important in these instances is knowing precisely how inaccurate their forecasts are. This can help them quantify their potential costs from differences in perceived and actual risk. Hedging strategies are found to effectively mitigate actual risks, while increased communication and building of supply chain relationships has been shown to mitigate the effects of perceived risk deviating from actual risk.

Future research on the effect of risk on forecast accuracy should explore the differences between perceived and actual risks in how they affect total supply chain costs. Current logistics and supply chain management research conflates these distinct types of uncertainty, and tend to focus on remedies that can be enacted by a single firm. Supply chains depend on the willing participation of multiple firms, each consisting of agents with varying risk preferences, so the effects of these differences on supply chain costs should be measured.

Production and Inventory Control Policy

Generally speaking, more accurate demand forecasts decrease production and inventory costs, regardless of control system. However, the effects of forecast accuracy

on costs depend on the type of control policies in place, and has been found to have Pareto returns.

Production and inventory replenishment policies that include small buffers and lot sizes, and who have little excess volume or varietal capacity will tend to be more sensitive to demand forecast errors and are likely to see the most benefit from increases in accuracy. Such policies are driven by achievable demand forecast accuracy, but also varying costs of production switching and setup, master production schedule (MPS) replanning, flexible production capacity, production volume capacity, labor, buffer and lead time inventory carrying, and shortage costs. While many of these costs offset each other depending on inventory or production policy, error has been found to significantly affect the choice of policy and the costs associated. Many significant examples exist regarding the choice of lot sizing policy, MPS freezing and replanning periodicity, buffer stock policy, and lead time replenishment policy.

Demand forecast error, along with material requirements planning (MRP) process error (Fildes and Kingsman 2011) and relevant cost tradeoffs between setup and inventory carrying costs (Ho and Ireland 2012) have been found to have a significant effect on the choice of lot sizing policy. Error can drive more frequent orders and more responsive lot size policies if setup costs are low, it can motivate less responsive lot sizing policies to take advantage of the effects of error aggregation is inventory holding costs are low (Ho and Ireland 2012). The interaction of lot sizing policy and demand forecast error have been found to have significant effects on overall production costs (Wemmerlöv 1985, Venkataraman and D'Itri 2001, Fildes and Kingsman 2011),

increasingly as MRP structures become more complex (Lee and Adam Jr. 1986). Though some studies have found cost savings from error reduction to be insensitive to lot sizing policy (Jeunet 2006), particularly in cases of capacity smoothing (Harl and Ritzman 1985). Lot sizing policy has also been found to offset the production instability effects of demand forecast error (Ho and Ireland 1993, 1998, 2012), except in cases of very high error (De Bodt and Van Wassenhove 1983).

Forecast errors significantly impact optimal MPS planning horizon, freezing proportion, freezing method (by period or order), and replanning periodicity depending on direction of error bias (Zhao and Lee 1993, Lin et al. 1994,Venkataraman and Nathan 1999, Xie et al. 2004). Demand forecast error effects are shown to dominate the effect of freezing proportion and replanning periodicity on MRP system costs (Yang and Jacobs 1999), particularly in push systems. In pull systems with frequent replanning and lower degrees of MPS freezing, demand forecast error has a smaller relative effect dependent on the relative costs of responsiveness and inventory holding (Masuchun et al. 2004). There are some cases, such as in multilevel, multiproduct production systems that experience hedging between product lines, that error is found to have linear effects on costs (Altendorfer et al. 2016). However, most research indicates inventory reductions and unit costs are not proportional to error reductions in production control environments (Doganis et al. 2008, Fildes and Kingsman 2011, Jeong 2011), and some production control policies are relatively more asymmetrically sensitive to bias than error (Xie et al. 2004, Sourirajan et al. 2008). When shortage costs are accounted for, slight positive bias (nonzero error) has been found to minimize costs (Biggs and Campion 1982, Lee and

Adam Jr. 1986, Xie et al. 2004) and improve delivery performance (Enns 2002), particularly when capacity is tight or shortage costs are high, but this has also been shown to increase MPS tardiness (Enns 2002).

Demand forecast error drives required levels of safety or buffer stock, but decreases in error provide Pareto returns on inventory cost savings (Zinn and Marmorstein 1990), and may better be accomplished through reductions in bias (Willemain et al. 2004). Buffer stock policy can also offsets forecast accuracy in reducing schedule instability (Ho and Ireland 1993, De Bodt and Van Wassenhove 1983, Enns 2002), though buffering is insufficient to completely control "nervousness" (Wemmerlöv 1985). Inventory costs are found to be more sensitive to increases in error when safety stock follows a constant cycle versus a constant stock policy, and when error is nonstationary (Campbell 1995), and stock performance depends on order frequency, lead time, and error distribution.

Logistics and supply chain management research also indicates significant effects from forecast accuracy on inventory control policy. Liao and Chang (2010) find interesting results under periodic review order-up-to (s,T) and continuous review fixed-order quantity (r,Q) varying lead time and error in an ant colony optimization. In (s,T) systems, longer lead time costs were positively correlated with error, whereas in an (r,Q) system shorter lead time costs were more positively correlated with error. This implies achievable error can be substituted for with appropriate inventory replenishment policy and lead time combinations, though in general longer lead times monotonically increase forecast error and overall inventory costs (Jeffery et al. 2008). Higher error associated

with longer lead times can also moderate the cost tradeoffs between quantity discounts and inventory carrying costs in an (r,Q) system (Kim et al. 2003). Tratar (2010) show that neither minimization of error nor bias under (s,T) replenishment minimizes inventory costs, and suggest instead a joint optimization. Though this could be improved upon by expanding to a more holistic measure like EVA. Huang et al. (2011) show that in a (s,T) system, order adjustments based on bias are more cost effective than those based on error. Ganeshan et al. (2001) show that in a four tier supply chain (distribution requirements planning system) for a chemical company, increased error is related to increases in cycle time costs (inventory carrying costs, obsolescence, warehousing) and reductions in service level and return on investment (due to stock-outs, on-time delivery performance and transportation costs). Inventory planning in closed loop systems may face difficulty achieving accurate forecasts due to the increased expense and reduced reliability of data in reverse channels (Kelle and Silver 1989). Finally, the effect of forecast accuracy on inventory costs can be shifted to other members of the supply chain through collaborative planning and control arrangements like vendor managed inventory (Kannan et al. 2013).

In general, this theme would imply "good enough" depends on the lot sizing policy, MPS freezing and replanning periodicity, buffer stock policy, and lead time replenishment policy in use. The appropriate choice of each policy and the effect of demand forecast error would then include a consideration of relevant cost tradeoffs. As with previous themes, work relating to control policies indicate Pareto returns on forecast accuracy investments, and that other metrics such as forecast bias may be more effective indicators of overall costs than forecast error.

Current research on this theme is limited to a production control policies in highly specific contexts. Future work should attempt to generalize the effect forecast accuracy on production control policies under multiple circumstances to determine whether product, industry or market factors are significant covariates. The majority of research regarding the effect of forecast accuracy on inventory control policy costs focuses on two control mechanisms. Future research in this area ought to include a greater focus on collaborative inventory models, hybrid inventory models, and inclusion of stochastic assumptions, as suggested by Williams and Tokar (2008).

Driver	*Implications*	*Proposed Future Research*
Error Amplification	"Good enough" depends on susceptibility to bullwhip, the relative import of the error amplification component of bullwhip costs, and feasibility of alternate means (other than error reduction) to combat these costs.	Must shift to more comprehensive diagnoses and normative responses.
Cost Tradeoffs	Desired accuracy may be lower depending on the complex tradeoffs, both pecuniary and non-pecuniary, within and between firms in a supply chain. To determine "good enough", firms must assess their relevant tradeoffs and prioritize accuracy improvement investments to those situations that promise the greatest returns.	Should focus on direct and indirect effects of demand forecast accuracy improvements on more holistic long term value measures such as EVA.
Supply Chain Integration	"Good enough" can change substantially based on cost of shared data, level of information asymmetry, trustworthiness of the shared data, for shared forecasts the accuracy of the downstream forecast, how collaborating shifts cost burdens and on the effectiveness and maturity of the collaborative relationship.	The different effects of various types of information to be shared and types of collaboration between supply chain principals have yet to be explored.
Supply Chain Flexibility	Flexible strategies have been shown to both increase achievable accuracy and reduce reliance on accuracy, potentially meaning lower desired accuracy. Flexible strategies affect "good enough" by both replacing and augmenting demand forecast accuracy.	Current research on flexibility has too great a focus on local operations costs, must focus on holistic metrics like EVA.
Manual Adjustments	Manual adjustments can be expensive to implement, but have been shown to increase achievable accuracy in a variety of circumstances. Practitioners must weigh costs of updates against the potential benefit to determine "good enough".	Effects of adjustments need to be studied over a greater variety of contexts to normatively prescribe action.
Risk	"Good enough" depends on both actual and perceived risk. Hedging against actual risk and increasing communication to limit the portion of perceived risk that differs from actual risk can replace higher required demand forecast accuracy.	Requires a greater understanding of the effect of differences between perceived and actual levels of risk over multiple supply chain principals.
Inventory and Control Policy	Inventory and production control policies introduce discontinuities in the cost tradeoff considerations of "good enough". Practitioners should consider the effects of such policies explicitly when measuring potential tradeoffs, and include other metrics besides accuracy (particularly bias).	Should examine more generalizable production conditions, include more stochastic assumptions, and focus on more complex collaborative and hybrid inventory models.

Table 17: Summary of Managerial Drivers of Demand Forecast Accuracy

Limitations

Our search was limited to journals with either a focus on logistics or supply chain management, or related journals that feature research that specifically mention implications of logistics or the supply chain. The term "logistics" only began to popularly replace "physical distribution management" in the 1970s (Farris 1997, Southern 2011). "Supply chain management" was coined in the 1980s, but did not gain prominence until the later 1990s (Cooper et al. 1997). By including these terms, and not more antiquated terms for similar concepts, we necessarily limited the search to more recent articles. We also focused on peer reviewed journals written in English. The limitation to English language articles reflects a limitation of the research team. Non-English articles likely hold merit and would significantly improve our exploration of the bounds of demand forecast accuracy, but our team did not possess the capabilities to assess this. Omitting non-peer reviewed material admittedly eliminated several theses, dissertations, books, and conference proceedings that certainly have merit and are closely related to our topic. However, as these works were not subject to rigorous peer review, our team had no verification of their rigor.

Conclusion

This review of logistics and supply chain management literature reveals a number of themes that have been found to shape the bounds to achievable and desirable forecast accuracy. Each theme has been explored and measured in academic research to varying degrees, and the expected effects have been measured over a wide range of conditions. For each theme, we provide a brief overview of extant academic work, relate what this

means for a demand manager, and recommend future directions for academic efforts to aid practice. For the practitioner, this provides some indication, supported empirically, analytically, or theoretically, of where to look when attempting to answer "How good is good enough?".

For technical drivers of demand forecast accuracy, the six themes indicate both positive and negative forces on primarily achievable forecast accuracy. Research on *forecastability* would suggest some demand patterns, characterized by high variation, intermittence, and other non-stationarity makes lower levels of forecast error unachievable. "Good enough" is lowered, and alternative mitigating investments, or metrics to prioritize efforts like RAE better serve managers. Extant research indicates increasing the *horizon* of a forecast also lowers achievable forecast accuracy, but that this effect can be offset by aggregation effects when forecasts are generated for alternate purposes. In these cases, managers must try to estimate what proportion of error comes from the horizon. *Overfitting and misspecification* also lower "good enough", as when forecasters develop extrapolative models, they must be wary of replicating past patterns too closely. This unfortunate ailment of all mathematical models implies that achievable accuracy is higher than desirable accuracy when modelers adapt explanatory models for use in prediction. In considering *tradeoffs of metrics* findings in the literature are that "good enough" may not matter unless the right type of deviation of prediction from reality is measured. Depending on circumstance, some metrics are more important to overall performance than others, and optimizing any one will cause others to suffer. Multiple measures should be used, and "good enough" for any one metric should be

lower than what is achievable. Research on level of *aggregation and hierarchy* indicate that increases in both of these dimensions drive "good enough" higher. The effect from hierarchy results from a concentration of resources. However, there is a risk from aggregating in misunderstanding the relative importance of sampling and specification error. Higher achievable accuracy through aggregation may also result in lower utility of a forecast. Finally, academic work regarding the effect of *data quality and availability* indicate that limits to these drivers will lower "good enough". Hard limits exist where improvements on quality or availability are not possible, but these can often be affected through investments in capabilities to collect information or incentives to share information between organizations.

Among managerial drivers of demand forecast accuracy, the seven themes indicate both positive and negative forces on primarily desirable forecast accuracy. *Error amplification* (or more commonly referred to as bullwhip) research demonstrates that the effect of demand forecast accuracy on bullwhip-related costs is often smaller than other factors. The literature also indicates signal amplification costs are only some of the relevant bullwhip-related costs, and that these are unequally shared based on a firm's position in a supply chain. This suggests other remedies such as bias or lead time reduction are more cost effective responses, and "good enough" may be lower. The most complex and heavily investigated theme, *cost tradeoffs*, reveals the interdependent direct, indirect and nonpecuniary costs associated with investment in additional forecast accuracy or costs associated with lower levels of accuracy. Work reviewed under this theme indicate the criticality for firms to accurately estimate their relevant costs, while

also revealing the incredible difficulty of generating cost (or value) functions for several possible alternatives, requiring voluntary coordination from supply chain members who bear costs and benefits unequally. Previous work indicates which costs will likely effect "good enough", but combination of effects depends on individual circumstances. Research on *supply chain integration* indicates "good enough" increases with increasing integration among upstream supply chain members, but "good enough" is lower with increasing integration downstream. This means that "good enough" becomes a function of position in a supply chain, and of how effectively the supply chain members can share relevant costs and benefits of information sharing. The literature would indicate *supply chain flexibility* has some mixed effects on "good enough". Flexible strategies simultaneously lower "good enough" by reducing reliance on forecast accuracy, and increase "good enough" by increasing achievable accuracy. *Manual adjustments* are found to generally increase "good enough", but these depends on the costs associated with gathering inputs and have diminishing returns. Previous investigations on *risk* indicate mixed effects on "good enough". Actual risk can entail both greater levels of demand uncertainty, which lowers "good enough", and greater cost vulnerability to uncertainty, which increases "good enough". If substantially different from actual risk, perceived risk introduces additional costs, but in this case the requirement for accurate knowledge of the level of error is affected more than the actual level of forecast error. Finally, work regarding the theme of *production and inventory control policy* indicates mixed effects on "good enough". While inventory and production cost savings from increasing forecast accuracy would indicate higher levels of "good enough", these have

Pareto returns, and choice of policy appears to have a greater effect on costs than forecast accuracy.

Within each theme, we also recommend areas for future research. While the recommendations for each theme were unique, some common deficiencies in extant literature were apparent. Future work in most themes must include more holistic measures of value, and include a greater scope of extra-firm considerations. Local costs do not motivate high level investments to change how a business operates, so measures of forecast accuracy must be associated with long term value to the greatest extent possible. Firms also do not exist independently of their networks of suppliers or customers, so considerations of how forecast accuracy at various levels affect relative costs and benefits between members of the supply chain must be included in future research. This research can also be extended by applying these identified themes in a case study evaluation of a firm or supply chain's demand management processes. By explicitly measuring these themes, it would provide a template for practitioners in evaluating their own conditions for achievable and desirable forecast accuracy.

Demand managers will still have to identify which of these themes hold relevance to their situation, and attempt to measure the relevant tradeoffs present in these themes in order to determine what level of demand forecast accuracy is "good enough". However, guided by research findings in these six technical and seven managerial themes, they are better equipped to assess the nuanced and complicated set of considerations and tradeoffs that shape the answer to "How good is good enough?".

Chapter 5: Conclusions

These three essays examine interrelated concerns in a critical supply chain input. Whether forecasts are used for replenishment, sales staffing, coordinating storage and transportation, or for sourcing and production, accuracy is critical to a supply chain's health. The three questions posed by the 4PL firm we partnered with on this research: "What is causing our replenishment forecast error?", "What predictive factors can help improve our demand forecast accuracy?", and "How good is good enough?" helped us to address three more general deficiencies in forecasting literature as it relates to logistics and the supply chain.

By identifying the effects of a previously unexplored driver of upstream (replenishment) forecast deviation and bias, we show that internally controllable factors can affect the performance of upstream replenishment. In our examination of inclusion of exogenous (weather) factors on demand forecast quality, we demonstrate the ability to harness readily available external information to improve prediction. Finally, our review of the bounding factors of forecast accuracy provide practitioners with the means to identify the levels of forecast accuracy that are achievable and desirable for their firm or supply chain.

In our first essay, we found that the franchise governance form does significantly affect both replenishment forecast deviation and bias. While the effect on deviation was consistent with previous research on the proclivities of franchisees (Brickley and Dark 1987, Norton 1988a, Bertagnoli 1989, Krueger 1991, Carney and Gedajlovic 1991, Kaufmann and Lafontaine 1994, Michael 2000, Yin and Zajac 2004, de Leeuw, Holweg and Williams 2011), our findings on the effect of governance form on bias would seem to contrast with previous indications from research on the operational effect of agency in organizations (Rubin 1978, Norton 1988a, Norton 1988b, Noren 1990, Krueger 1991). These results suggest that governance form does indeed significantly drive behavior potentially misaligned with parent firm incentives, but the explanation for these differences may be more nuanced than previously identified.

In addition to these main findings, we developed a novel method to explore extremely large datasets in order to quantify the heterogeneities of the effect of governance form on replenishment forecast deviation and bias. The technique we call *HPD* permits the parsimonious identification of regional, temporal and product category differences in the effect so that firm resources can be effectively targeted. This is important, as firms are gathering data faster than they can effectively utilize it, and are urgently seeking means to leverage this resource into a competitive advantage.

Given the decentralized structure of the franchising governance form, there is a continued need to research factors that drive differences in post-contractual performance. Knowing these factors can guide firms in their efforts to align their mix of governance form with their overall strategy to minimizing form-specific residual loss and maximizing

the advantages of the plural form (Bradach 1997, Yin and Zajac 2004, Barthélemy 2008). If deviation or bias related to governance form is beneficial, the individual differences can be benchmarked (Bradach 1997, Yin and Zajac 2004). If not, firms can us relational governance to effect change in outlets where they have little coercive power (Bradach 1997, Paik and Choi 2007, Cochet et al. 2008).

In our second essay, we find multiple predicted weather factors that can significantly improve demand forecast accuracy. We found that weather forecasts for high and low temperatures, as well as for thunderstorms significantly improved demand forecast accuracy. We also found that other predicted weather such as wind speed, cloud cover, rain, snow and overall precipitation largely did not improve demand forecast accuracy, and in some cases made it worse. Effects were similar, if slightly more positive for models that utilized perfect weather prediction (observed rather than predicted weather). These effects differed by accuracy measure, product category, and weather region.

These mixed results indicate that inclusion of exogenous information into demand forecasts can prove a difficult and complicated matter, despite the promise of greater precision for operational planning (Bertrand and Sinclair-Desgagné 2011, Nikolopoulos and Fildes 2013, Steinker et al. 2016). This means that potential differences in the effect of inclusion of these exogenous factors depends not only on the identification of a significant weather effect, but the correct specification of that effect, which can be nonlinear and heterogeneous across a number of dimensions. There is additionally the potential for confounding present in aggregation of any sort of effect for use in demand

forecasting. The reliability of the exogenous factors themselves are also a concern worth monitoring, as each weather prediction differs in quality based on forecast horizon, specific weather phenomena, the sensitivity of weather sensors, and their distance from a relevant demand point. Finally, the type of measurement used for demand forecast error can affect the perceived benefit of including external weather predictions. The fact that most of the observed demand accuracy improvement was in percent error measures, and not in absolute measures, indicates that inclusion of external weather predictions increased the responsiveness of demand forecasts.

This supports previous findings that predicted weather can significantly improve demand forecasts (Nikolopoulos and Fildes 2013, Steinker et al. 2016), but extends these finding s to a new industry, and over a broader set of regional and product-based circumstances. It also demonstrates some of the difficulties inherent in inclusion of uncertain exogenous information in demand forecasts.

Our third essay discusses in detail the current state of logistics and supply chain research on the bound of forecast accuracy. Through our systematic literature review, we identify six technical and seven managerial themes found to drive differential levels of both achievable and desirable demand forecast accuracy. For each theme, we comprehensively described the current state of logistics and supply chain research, the implications of research for practitioners, and potential directions for future work. Our review indicates the numerous factors that can shape the levels of demand forecast accuracy are highly specific to the context, goals and capabilities of a firm or a supply chain. Beyond identifying the state and future direction of research, our themes provide a

guide for practitioners in assessing their own demand planning situation to determine the relevant factors that drive higher and in some cases lower forecast accuracy.

In addressing these three specific concerns from demand planners in a large 4PL, we expand the academic understanding in three interrelated topics in forecast accuracy as it related to logistics and the supply chain. Our results should drive future research on factors affecting upstream replenishment forecast accuracy, factors that may improve downstream demand forecast accuracy, and themes that dictate achievable and desirable levels of demand prediction accuracy.

References

Agrawal, S., Sengupta, R.N. and Shanker, K., 2009. Impact of information sharing and lead time on bullwhip effect and on-hand inventory. *European Journal of Operational Research*, 192, pp.576–593.

Ahlburg, D.A., 1992. A commentary on error measures: Error measures and the choice of a forecast method. *International Journal of Forecasting*, 8, pp.99–100.

Ali, M.M., Boylan, J.E. and Syntetos, A.A., 2012. Forecast errors and inventory performance under forecast information sharing. *International Journal of Forecasting*, 28(4), pp.830–841. Available at: http://dx.doi.org/10.1016/j.ijforecast.2010.08.003.

Altendorfer, K., Felberbauer, T. and Jodlbauer, H., 2016. Effects of forecast errors on optimal utilisation in aggregate production planning with stochastic customer demand. *International Journal of Production Research*, 54(12), pp.3718–3735. Available at: http://dx.doi.org/10.1080/00207543.2016.1162918.

Amornpetchkul, T., Duenyas, I. and Şahin, Ö., 2015. Mechanisms to Induce Buyer Forecasting: Do Suppliers Always Benefit from Better Forecasting? *Production and Operations Management*, 24(11), pp.1724–1749.

AMS (American Meteorological Society)., 2015. Weather Analysis and Forecasting: An Information Statement of the American Meteorological Society. AMS Council.

Armstrong, J.S., 2002. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. (J.S. Armstrong, Ed.), *International Series in Operations Research & Management Science* (Vol. 18). New York: Kluwer Academic Publishers.

Armstrong, J.S. and Collopy, F., 1992. Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons. *International Journal of Forecasting*, 8(1), pp.69–80. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=662701.

Armstrong, J.S. and Fildes, R., 1995. Correspondence on the Selection of Error Measures for Comparisons among Forecasting Methods. *Journal of Forecasting*, 14(August 1994), pp.67–71.

Arrow, K.J., 1965. Aspects of the theory of risk-bearing. Yrjö Jahnssonin Säätiö.

Arunraj, N.S. and Ahrens, D., 2016. Estimation of non-catastrophic weather impacts for retail industry. *International Journal of Retail & Distribution Management*, 44(7), pp.731–753. Available at: http://dx.doi.org/10.1108/IJRDM-07-2015-0101%5Cnhttp://dx.doi.org/10.1108/.

Arunraj, N.S., Ahrens, D. and Fernandes, M., 2016. Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry. *International Journal of Operations Research and Information Systems*, 7(2), pp.1–21.

Auffhammer, M. and Mansur, E.T., 2014. Measuring climatic impacts on energy consumption: A review of the empirical literature. *Energy Economics*, 46, pp.522–530. Available at: http://dx.doi.org/10.1016/j.eneco.2014.04.017.

Babai, M.Z., Ali, M.M., Boylan, J.E. and Syntetos, A.A., 2013. Forecasting and inventory performance in a two-stage supply chain with ARIMA (0,1,1) demand: Theory and empirical analysis. *International Journal of Production Economics*, 143(2), pp.463–471. Available at: http://dx.doi.org/10.1016/j.ijpe.2011.09.004.

Babai, M.Z., Syntetos, A.A. and Teunter, R., 2014. Intermittent demand forecasting: An empirical study on accuracy and the risk of obsolescence. *International Journal of Production Economics*, 157, pp.212–219. Available at: http://dx.doi.org/10.1016/j.ijpe.2014.08.019.

Bahng, Y. and Kincade, D. H., 2012. The relationship between temperature and sales. *International Journal of Retail & Distribution Management*, 40(6), 410–426. https://doi.org/10.1108/09590551211230232.

Banker, S., 2009. "Nestle Waters and Weather-Driven Demand." *Logistics Viewpoints*. ARC Advisory Group. https://logisticsviewpoints.com/2009/07/09/nestle-and-weather-driven-demand/.

Barman, S., Tersine, R.J. and Burch, E.E., 1990. Performance evaluation of the LDR and the PSH with forecast errors. *Journal of Operations Management*, 9(4), pp.481–499.

Barthélemy, J. 2008. "Opportunism, Knowledge, and the Performance of Franchise Chains". *Strat. Mgmt. J.* 29 (13): 1451-1463. Doi:10.1002/smj.719.

Bassi, A., Colacito, R. and Fulghieri, P., 2013. 'O Sole Mio: An Experimental Analysis of Weather and Risk Attitudes in Financial Decisions. *The Review of Financial Studies*, (919), pp.1–29.

Bertagnoli, L., 1989. "McDonald's: Company of the Quarter Century". *Restaurants and Institutions*, 32.

Bertrand, J. L., Brusset, X., and Fortin, M., 2015. Assessing and hedging the cost of unseasonal weather: Case of the apparel sector. *European Journal of Operational Research*, 244(1), 261–276. https://doi.org/10.1016/j.ejor.2015.01.012.

Bertrand, J. and Sinclair-Desgagné, B., 2011. Environmental Risks and Financial Markets: A Two-Way Street. In P. Bansal and A. J. Hoffman (Eds.), *The Oxford Handbook of Business and the Natural Environment* (pp. 1–26). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199584451.003.0026.

Bian, W., Shang, J. and Zhang, J., 2016. Two-way information sharing under supply chain competition. *International Journal of Production Economics*, 178, pp.82–94. Available at: http://dx.doi.org/10.1016/j.ijpe.2016.04.025.

Biggs, J.R. and Campion, W.M., 1982. The Effect and Cost of Forecast Error Bias for Multiple-Stage Production-Inventory Systems. *Decision Sciences*, 13, pp.570–584.

Bitner, M.J., 2017. Servicescapes: The Impact of Physical Surroundings on Customers and Employees. *Journal of Marketing*, 56(2), pp.57–71.

Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M., 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.

Bowersox, D.J., Closs, D.J., Cooper, M.B. and Bowersox, J.C., 2013. *Supply chain logistics management* (4th ed). New York, NY: McGraw-Hill.

Boylan, J.E. and Syntetos, A.A., 2006. Accuracy and Accuracy Implication Metrics for Intermittent Demand. *Foresight: International Journal of Applied Forecasting*, (4), pp.39–42.

Bradach, J.L., 1997. Using the Plural Form in the Management of Restaurant Chains. *Administrative Science Quarterly* 42 (2): 276. Doi: 10.2307/2393921.

Bratina, D. and Faganel, A., 2008. Forecasting the Primary Demand for a Beer Brand Using Time Series Analysis. *Organizacija*, 41(3), pp.116–124.

Breiter, A. and Huchzermeier, A., 2015. Promotion planning and supply chain contracting in a high-low pricing environment. *Production and Operations Management*, 24(2), pp.219–236.

Brenner, H. and Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine*, 16(9), pp.981–991.

Brickley, J.A. and Dark, F.H., 1987. The Choice of Organizational Form the Case of Franchising. *Journal of Financial Economics* 18 (2): 401-420. Doi:10.1016/0304-405x(87)90046-8.

Buchholz, W., 1962. *Planning a Computer System: Project Stretch*.

Bujisic, M., Bogicevic, V. and Parsa, H.G., 2016. The effect of weather factors on restaurant sales. *Journal of Foodservice Business Research*, 0(0), pp.1–21. Available at: http://dx.doi.org/10.1080/15378020.2016.1209723.

Burton, M., 2017. South Australia suffers another weather-induced blackout. *Reuters*. http://in.reuters.com/article/australia-electricity-outages-idINKBN15N2QY.

Busse, M.R., Pope, D.G., Pope, J.C. and Silva-Risson, J., 2012. Projection bias in the housing and car markets. *Unpublished manuscript.*

Cachon, G.P. and Fisher, M., 2000. Supply Chain Inventory Management and the Value of Shared Information. *Management Science*, 46(8), pp.1032–1048.

Cachon, G. and Lariviere, M., 2001. Contracting to Assure Supply: How to Share Demand Forecasts in a Supply Chain. *Management Science*, 47(5), pp.629–646.

Camargo, M.B.P. and Hubbard, K.G., 1999. Spatial and temporal variability of daily weather variables in sub-humid and semi-arid areas of the United States high plains. *Agricultural and Forest Meteorology*, 93(2), 141–148. https://doi.org/10.1016/S0168-1923(98)00122-1.

Campbell, G.M., 1995. Establishing safety stocks for master production schedules. *Production Planning & Control*, 6(5), pp.404–412. Available at: http://www.tandfonline.com/doi/abs/10.1080/09537289508930297.

Caniato, F., Kalchschmidt, M., Ronchi, S., Verganti, R. and Zotteri, G., 2005. Clustering customers to forecast demand. *Production Planning & Control*, 16(1), pp.32-43.

Canitz, H., 2016. Overcoming Barriers to Improving Forecast Capabilities. *Foresight: International Journal of Applied Forecasting*, Spring, pp.26–35.

Carney, M. and Gedajlovic, E. 1991. "Vertical Integration in Franchise Systems: Agency Theory and Resource Explanations". *Strategic Management Journal.* 12 (8): 607-629. Doi:10.1002/smj.4250120804.

Carter, C.R. and Liane Easton, P., 2011. Sustainable supply chain management: evolution and future directions. *International Journal of Physical Distribution & Logistics Management*, 41(1), pp.46-62.

Caves, R.E. and Murphy II, W.F.. 1976. Franchising: Firms, Markets, and Intangible Assets. *Southern Economic Journal* 42 (4): 572. Doi:10.2307/1056250.

Cawthorn, C., 1998. Weather as a strategic element in demand chain planning. *The Journal of Business Forecasting Methods & Systems*, 17(3), pp.18–21.

Census Bureau., 2012. Franchise Status for Selected Industries and States. Part of: *Core Business Statistics Series, 2012*. Data Set: *2012 Economic Census of the United States*. Available at American FactFinder, http://factfinder.census.gov; Accessed: 1/3/17.

Chang, S.Y. and Yeh, T.Y., 2012. Applying TLFs to design the exception thresholds in collaborative forecasting. *International Journal of Production Research*, 50(7), pp.1932–1941. Available at: http://www.tandfonline.com/doi/abs/10.1080/00207543.2011.564669.

Chatfield, C., 1992. A commentary on error measures. *International Journal of Forecasting*, 8, pp.100–102.

Chen, A. and Blue, J., 2010. Performance analysis of demand planning approaches for aggregating, forecasting and disaggregating interrelated demands. *International Journal of Production Economics*, 128(2), pp.586–602. Available at: http://dx.doi.org/10.1016/j.ijpe.2010.07.006.

Chen, F., 1999. Decentralized Supply Chains Subject to Information Delays. *Management Science*, 45(8), pp.1076–1090.

Chen, Y. and Xiao, W., 2012. Impact of Reseller's Forecasting Accuracy on Channel Member Performance. *Production and Operations Management*, 21(6), pp.1075–1089.

Choi, C., Kim, E. and Kim, C., 2011. A Way of Managing Weather Risks Considering Apparel Consumer Behaviors. *Working Paper*. Available at: http://ssrn.com/abstract=1909185.

Chopra, S. and Sodhi, M.S., 2004. Managing risk to avoid supply-chain breakdown. *MIT Sloan Management Review*, 46(1), p.53.

Christopher, M., 2000. The agile supply chain: competing in volatile markets. *Industrial Marketing Management*, 29(1), pp.37-44.

Chung, C., Niu, S.C. and Sriskandarajah, C., 2012. A sales forecast model for short-life-cycle products: New releases at blockbuster. *Production and Operations Management*, 21(5), pp.851–873.

Clark, A.R., 1998. Batch sequencing and sizing with regular varying demand. *Production Planning & Control*, 9(3), pp.260-266.

Clark, A.R., 2005. Rolling horizon heuristics for production planning and set-up scheduling with backlogs and error-prone demand forecasts. *Production Planning & Control*, 16(1), pp.81-97.

Clemen, R.T., 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4), pp.559-583.

Clemen, R.T. and Winkler, R.L., 1985. Limits for the Precision and Value of Information from Dependent Sources. *Operations Research*, 33(2), pp.427–443.

Cochet, O., Dormann, J. and Ehrmann, T., 2008. Capitalizing On Franchisee Autonomy: Relational Forms of Governance as Controls in Idiosyncratic Franchise Dyads*. *Journal of Small Business Management* 46 (1): 50-72. Doi:10.1111/j.1540-627x.2007.00231.x.

Cohen, J., Cohen, P., West, S.G. and Aiken, L.S., 2013. *Applied Multiple Regression/Correlation Analysis For The Behavioral Sciences*. Routledge.

Combs, J.G. and Ketchen, D.J., 1999. Can Capital Scarcity Help Agency Theory Explain Franchising? Revisiting the Capital Scarcity Hypothesis. *Academy of Management Journal* 42 (2): 196-207. Doi:10.2307/257092.

Combs, J.G. and Ketchen, D.J., 2003. Why Do Firms Use Franchising As An Entrepreneurial Strategy?: A Meta-Analysis. *Journal of Management* 29 (3): 443-465. Doi:10.1016/s0149-2063_03_00019-9.

Combs, J.G., Ketchen, D.J., Shook, C.L. and Short, J.C., 2010. Antecedents and Consequences of Franchising: Past Accomplishments and Future Challenges. *Journal of Management* 37 (1): 99-126. Doi:10.1177/0149206310386963.

Considine, T.J., 2000. The impacts of weather variations on energy demand and carbon emissions. *Resource and Energy Economics*, 22, pp.295–314.

Cooper, M.C., Lambert, D.M. and Pagh, J.D., 1997. Supply chain management: more than a new name for logistics. *The International Journal of Logistics Management*, 8(1), pp.1-14.

Coronado Mondragon, A.E., Lalwani, C. and Coronado Mondragon, C.E., 2011. Measures for auditing performance and integration in closed-loop supply chains. *Supply Chain Management: An International Journal*, 16(1), pp.43–56. Available at: http://www.emeraldinsight.com/doi/10.1108/13598541111103494.

Cotteleer, M.J. and Wan, X., 2016. Does The Starting Point Matter? The Literature-Driven and The Phenomenon-Driven Approaches of Using Corporate Archival Data In Academic Research. *Journal of Business Logistics*, *37*(1), Pp.26-33.

Crowther, M.A. and Cook, D.J., 2007. Trials and tribulations of systematic reviews and meta-analyses. *ASH Education Program Book*, 2007(1), pp.493-497.

Cui, R., Allon, G., Bassamboo, A. and Van Mieghem, J.A., 2015. Information Sharing in Supply Chains: An Empirical and Theoretical Valuation Information. *Management Science*, 61(11).

Cunningham, M.R., 1979. Weather, Mood, and Helping Behavior: Quasi Experiments with the Sunshine Samaritan. *Journal of Personality and Social Psychology*, 37(11), pp.1947–1956.

Dalrymple, D.J., 1975. Sales Forecasting Methods and Accuracy. *Business Horizons*, (December).

Darlington, R.B. and Hayes, A.F., 1990. *Regression and Linear Models*. New York: McGraw-Hill.

Davis, L.B. et al., 2016. Analysis and prediction of food donation behavior for a domestic hunger relief organization. *International Journal of Production Economics*, 182, pp.26–37. Available at: http://dx.doi.org/10.1016/j.ijpe.2016.07.020.

De Bodt, M.A. and Van Wassenhove, L.N., 1983. Cost Increases Due to Demand Uncertainty in MRP Lot Sizing. *Decision Sciences*, 14, pp.345–362.

De Brito, M.P. and van der Laan, E.A., 2009. Inventory control with product returns: The impact of imperfect information. *European Journal of Operational Research*, 194(1), pp.105–121. Available at: http://dx.doi.org/10.1016/j.ejor.2007.11.063.

De Leeuw, S., Holweg, M. and Williams, G. 2011. The Impact of Decentralised Control on Firm-Level Inventory. *International Journal of Physical Distribution & Logistics Management* 41 (5): 435-456. Doi:10.1108/09600031111138817.

Denyer, D. and Neely, A., 2004. Introduction to special issue: innovation and productivity performance in the UK. *International Journal of Management Reviews*, 5(3-4), pp.131-135.

Dillow, C., 2011. In Brazil, an Explosion in Computing Power is Revolutionizing Weather Prediction. *Popular Science*. Available at: http://www.popsci.com/science/article/2011-05/better-weather-explosion-computing-power-fueling-weather-modeling-revolution.

Divakar, S., Ratchford, B. and Shankar, V., 2005. CHAN4CAST: A Multichannel, Multiregion Sales Forecasting Model and Decision Support System for Consumer Packaged Goods. *Marketing Science*, 24(3), pp.334–350. Available at: http://mktsci.journal.informs.org/content/24/3/334.abstract.

Doering, T. and Suresh, N.C., 2016. Forecasting and Performance: Conceptualizing Forecasting Management Competence as a Higher-Order Construct. *Journal of Supply Chain Management*, 52(4), pp.77–91.

Doganis, P., Aggelogiannaki, E. and Sarimveis, H., 2008. A combined model predictive control and time series forecasting framework for production-inventory systems.

International Journal of Production Research, 46(24), pp.6841–6853. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-55349105856&partnerID=tZOtx3y1.

Donovan, R.J., 1994. Store atmosphere and purchasing behavior. *Journal of Retailing*, 70(3), pp.283–294.

Durach, C.F., Kembro, J. and Wieland, A., 2017. Systematic Literature Reviews: Addressing the Ontological and Epistemological Idiosyncrasies of Supply Chain Management Research. *Journal of Supply Chain Management*. Forthcoming.

Dutton, J. A., 2002. Opportunites and Priorities in a New Era for Weather and Climate Services. *American Meteorological Society*, (May), 1303–1311.

Ebert, R.J. and Lee, T.S., 1995. Production loss functions and subjective assessments of forecast errors: untapped sources for effective master production scheduling. *International Journal of Production Research*, 33(1), pp.137–159.

Ebrahim-Khanjari, N., Hopp, W. and Iravani, S.M.R., 2012. Trust and information sharing in supply chains. *Production and Operations Management*, 21(3), pp.444–464.

Eksoz, C., Mansouri, S.A. and Bourlakis, M., 2014. Collaborative forecasting in the food supply chain: A conceptual framework. *International Journal of Production Economics*, 158, pp.120–135. Available at: http://dx.doi.org/10.1016/j.ijpe.2014.07.031.

Ellram, L.M. and Cooper, M.C., 2014. Supply chain management: It's all about the journey, not the destination. *Journal of Supply Chain Management*, 50(1), pp.8-20.

Enns, S.T., 2002. MRP performance effects due to forecast bias and demand uncertainty. European *Journal of Operational Research*, 138, pp.87–102.

Eroglu, C. and Croxton, K.L., 2010. Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting*, 26(1), pp.116–133. Available at: http://dx.doi.org/10.1016/j.ijforecast.2009.02.005.

Eroglu, C. and Knemeyer, A.M., 2010. Exploring the Potential Effects of Forecaster Motivational Orientation and Gender on Judgmental Adjustments of Statistical Forecasts. *Journal of Business Logistics*, 31(1), pp.179–195.

Fama, E.F. and Jensen, M.C., 1983. Separation of Ownership and Control. *Journal of Law and Economics*, 301-325.

Farris, M.T., 1997. Evolution of academic concerns with transportation and logistics. *Transportation Journal*, 37(1), pp.42-50.

Fenimore, C., 2017. U.S. Climate Divisions. *National Climatic Data Center, NOAA*, www.ncdc.noaa.gov/monitoring-references/maps/us-climate-divisions.php.

Fildes, R., 1992. The evaluation of extrapolative forecasting methods. *International Journal of Forecasting*, 8, pp.81–98.

Fildes, R., Goodwin, P. and Onkal, D., 2015. Information use in supply chain forecasting. *(Dept. Management Science Working Paper 2015:2).* Lancaster University.

Fildes, R. and Kingsman, B., 2011. Incorporating demand uncertainty and forecast error in supply chain planning models. *Journal of the Operational Research Society*, 62(3), pp.483–500. Available at: http://link.springer.com/10.1057/jors.2010.40.

Fildes, R., Nikolopoulos, K., Crone, S.F. and Syntetos, A.A., 2008. Forecasting and Operational Research: A Review. *Journal of the Operational Research Society.* 59 (9): 1150-1172. Doi:10.1057/palgrave.jors.2602597.

Fildes, R., Goodwin, P., Lawrence, M. and Nikolopoulos, K., 2009. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25(1), pp.3–23. Available at: http://dx.doi.org/10.1016/j.ijforecast.2008.11.010.

Fildes, R. and Petropoulos, F., 2015. Improving Forecast Quality in Practice. *Foresight: International Journal of Applied Forecasting*, Winter, pp.5–13.

Fisher, M. and Raman, A., 1996. Reducing the Cost of Demand Uncertainty through Accurate Response to Early Sales. *Operations Research*, 44(1), pp.87–99.

Fliedner, E.B. and Lawrence, B., 1995. Forecasting system parent group formation: An empirical application of cluster analysis. *Journal of Operations Management*, 12, pp.119–130.

Floehr, E., 2015. Analysis of One- to Nine-Day-Out High Temperature Forecasts for U.S., Europe, and Asia-Pacific (Calendar Year 2014).

Flores, B.E., Olson, D.L. and Pearce, S.L., 1993. Cost and Accuracy Measures in Forecasting Method Selection: a Physical Distribution Example. *International Journal of Production Research*, 31(1), pp.139–160. Available at: http://cat.inist.fr/?aModele=afficheN&cpsidt=4228222%5Cnpapers3://publication/uuid/21EFC91D-401B-4863-A369-7591AA4D0113.

Flores, B.E. and Wichern, D.W., 2005. Evaluating Forecasts: A Look at Aggregate Bias and Accuracy Measures. *Journal of Forecasting*, 24, pp.433–451.

Forrester, J.W., 1958. Industrial dynamics-a major breakthrough for decision makers. *Harvard Business Review*, 36(4), p.37.

Fox, J. and Weisberg, S., 2010. *An R Companion to Applied Regression*. Sage.

Fustier, J., 2011. PepsiCo, peu convaincu par la fiabilité des prévisions météo. *Supply Chain Magazine*, Juillet-Ao.

Ganeshan, R., Boone, T. and Stenger, A.J., 2001. The impact of inventory and flow planning parameters on supply chain performance: An exploratory study. *International Journal of Production Economics*, 71(18).

Gardner, E.S., 1990. Evaluating forecast performance in an inventory control system. *Management Science* 36, no. 4: 490-499.

Gardner, E.S. and Acar, Y., 2016. The forecastability quotient reconsidered. *International Journal of Forecasting*, 32(4), pp.1208-1211.

Gardner, M.P., 1985. Mood States and Consumer Behavior: A Critical Review. *Journal of Consumer Research*, 12(Dec), p.281=300.

Gaur, V., Kesavan, S., Raman, A. and Fisher, M.L., 2007. Estimating Demand Uncertainty Using Judgmental Forecasts. *Manufacturing & Service Operations Management*, 9(4), pp.480–491.

Georgiadis, P. and Vlachos, D., 2006. The Impact of Product Lifecycle on Capacity Planning of Closed-Loop Supply Chains with Remanufacturing. *Production and Operations Management*, 15(4), pp.514–527. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1937-5956.2006.tb00160.x/abstract.

Goldstein, K.M., 1972. Weather, mood, and internal–external control. *Perceptual and Motor Skills*. 35 (August), 786.

Gönül, M.S. and Goodwin, P., 2012. Why Should I Trust Your Forecasts? *Foresight: International Journal of Applied Forecasting*, pp.5–10.

Goodwin, P., 2011. High on Complexity, Low on Evidence: Are Advanced Forecasting Methods Always as Good as They Seem? *Foresight: International Journal of Applied Forecasting*, pp.10–13.

Graman, G.A. and Sanders, N.R., 2009. Modelling the tradeoff between postponement capacity and forecast accuracy. *Production Planning and Control*, 20(3), pp.206-215.

Grimm, C., Knemeyer, M., Polyviou, M. and Ren, X., 2015. Supply chain management research in management journals: A review of recent literature (2004-2013). *International Journal of Physical Distribution & Logistics Management*, 45(5), pp.404-458.

Gu, Q., Li, Z. and Han, J., 2012. Generalized Fisher Score For Feature Selection. *Arxiv Preprint Arxiv:1202.3725*.

Guo, Z., Fang, F. and Whinston, A.B., 2006. Supply chain information sharing in a macro prediction market. *Decision Support Systems*, 42(3), pp.1944–1958.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L., 2006. *Multivariate Data Analysis*. Vol. 6. Upper Saddle River, NJ: Pearson Prentice Hall.

Hamjah, M.A., 2014. Temperature and Rainfall Effects on Spice Crops Production and Forecasting the Production in Bangladesh : An Application of Box- Jenkins ARIMAX Model. *Mathematical Theory and Modeling*. Shahjalal University of Science and Technology.

Han, J., Pei, J. and Kamber, M., 2011. *Data Mining: Concepts and Techniques*. Elsevier.

Hançerlioğulları, G., Şen, A. and Aktunç, E.A., 2016. Demand uncertainty and inventory turnover performance: An empirical analysis of the US retail industry. *International Journal of Physical Distribution & Logistics Management*, 46(6/7), pp.681–708.

Hand, D., Mannila H. and Smyth P., 2001. *Principles of Data Mining*. MIT Press. Vol. 30. Doi:10.2165/00002018-200730070-00010.

Hardy, Q., 2015. IBM to Acquire the Weather Company. *The New York Times*. https://www.nytimes.com/2015/10/29/technology/ibm-to-acquire-the-weather-company.html?_r=1.

Harl, J.E. and Ritzman, L.P., 1985. A heuristic algorithm for capacity sensitive requirements planning. *Journal of Operations Management*, 5(3), pp.309–326.

Hartzel, K.S. and Wood, C.A., 2017. Factors that affect the improvement of demand forecast accuracy through point-of-sale reporting. *European Journal of Operational Research*, 260(1), pp.171–182. Available at: http://dx.doi.org/10.1016/j.ejor.2016.11.047.

Hatzakis, E.D., Nair, S.K. and Pinedo, M.L., 2010. Operations in financial services – An overview. *Production and Operations Management*, 19(6), pp.633–664. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-79952036300&partnerID=40&md5=da6ff8eec20732321daf7d1b5fa2cd5a.

Heckman, J.J., 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. In *Annals of Economic and Social Measurement*, Volume 5, Number 4 (Pp. 475-492). NBER.

Henningsen, A., 2010. Estimating Censored Regression Models In R Using The Censreg Package. *University of Copenhagen*.

Hijmans, R.J., Williams, E. and Vennes, C., 2016. Package 'geosphere': *Spherical Trigonometry*.

Hirshleifer, D. and Shumway, T., 2017. American Finance Association Good Day Sunshine: Stock Returns and the Weather. *The Journal of Finance*, 58(3), pp.1009–1032.

Ho, C.J. and Ireland, T.C., 1993. A diagnostic analysis of the impact of forecast errors on production planning via MRP system nervousness. *Production Planning & Control*, 4(4), p.311. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=5793096&lang=es&site=ehost-live.

Ho, C.J. and Ireland, T.C., 1998. Correlating MRP system nervousness with forecast errors. *International Journal of Production Research* ISSN: 36(8), pp.2285–2299.

Ho, C. and Ireland, T.C., 2012. Mitigating forecast errors by lot-sizing rules in ERP-controlled manufacturing systems. *International Journal of Production Research*, 50(11), pp.3080–3094.

Holmström, J., Korhonen, H., Laiho, A. and Hartiala, H., 2006. Managing product introductions across the supply chain: findings from a development project. *Supply Chain Management: An International Journal*, 11(2), pp.121–130. Available at: http://www.emeraldinsight.com/doi/10.1108/13598540610652519.

Hoover's Inc., 2016. Fast-Food & Quick-Service Restaurants. Retrieved August 18, 2016 From *Hoover's Database*.

Horn, J.L., 1965. A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, 30(2), Pp.179-185.

Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied Logistic Regression* (Vol. 398). John Wiley & Sons.

Hosoda, T. and Disney, S.M., 2006. On variance amplification in a three-echelon supply chain with minimum mean square error forecasting. *Omega: International Journal of Management Science*, 34(4), pp.344–358.

Hosoda, T. and Disney, S.M., 2009. Impact of market demand mis-specification on a two-level supply chain. *International Journal of Production Economics*, 121(2), pp.739–751. Available at: http://dx.doi.org/10.1016/j.ijpe.2009.04.024.

Howells, T. and Morgan, E., 2017. Gross Domestic Product by Industry: Fourth Quarter and Annual 2016. *US Department of Commerce, Bureau of Economic Analysis*.

Hu, Q.S. and Skaggs, K., 2009. Accuracy of 6-10 Day Precipitation Forecasts and Its Improvement in the Past Six Years. In *7th NOAA Annual Climate Prediction Application Science Workshop*. Pp. 1–2.

Huang, L.T., Hsieh, I.C. and Farn, C.K., 2011. On ordering adjustment policy under rolling forecast in supply chain planning. *Computers and Industrial Engineering*, 60(3), pp.397–410. Available at: http://dx.doi.org/10.1016/j.cie.2010.07.018.

Huang, Y.S., Hung, J.S. and Ho, J.W., 2017. A study on information sharing for supply chains with multiple suppliers. *Computers & Industrial Engineering*, 104, pp.114–123. Available at: http://dx.doi.org/10.1016/j.cie.2016.12.014.

Huntemann, T.L., Rudack, D.E., and Ruth, D.P., 2014. Forty Years of NWS Verification: Past Performance and Future Advances.

Hur, D., Mabert, V.A. and Bretthauer, K.M., 2009. Real-Time Work Schedule Adjustment Decisions: An Investigation and Evaluation. *Production and Operations Management*, 13(4), pp.322–339. Available at: http://doi.wiley.com/10.1111/j.1937-5956.2004.tb00221.x.

Hyndman, R.J., 2006. Another Look at Forecast Accuracy Metrics for Intermittent Demand. *Foresight: International Journal of Applied Forecasting*, 4(June), pp.43–46.

Hyndman, R.J. and Athanasopoulos, G., 2014. *Forecasting: principles and practice*. Otexts.

Hyndman, R.J. and Khandakar, Y., 2008. Automatic time series forecasting: The forecast package for R., *Journal of Statistical Software*, 26(3).

Hyndman, R.J. and Koehler, A.B., 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(1), pp.679–688.

Hyndman, R.J. and Kostenko, A. V, 2007. Minimum Sample Size Requirements for Seasonal Forecasting Models. *Foresight: The International Journal of Applied Forecasting*, (6), pp.12–15.

Hyndman, R.J., O'Hara-Wild, M., Bergmeir, C., Razbash, S. and Wang, E., 2017. Package 'forecast': Forecasting Functions for Time Series and Linear Models.

IGD (The Institute of Grocery Distribution)., 2009. Case Study: Nestlé Waters – Demand Planning Weather Project. *The Institute of Grocery Distribution (GB).* https://www.igd.com/Research/Supply-chain/Waste-prevention/Six-to-fix-to-prevent-waste/Forecast/Nestle-Waters---Demand-planning-weather-project/.

Jain, C., 2001. Forecasting Practices in Corporate America. *The Journal of Business Forecasting*, *20*(1).

Janis, M.J., Hubbard, K.G. and Redmond, K.T., 2004. Station density strategy for monitoring long term climatic change in the contiguous United States. *Journal of Climate*, 17(2001), pp.151–163. Available at: http://climate.geog.udel.edu/~climate/publication_html/Pdf/JHR_JClim_04.pdf.

Jeffery, M.M. et al., 2008. Determining a cost-effective customer service level. *Supply Chain Management: An International Journal*, 13(3), pp.225–232.

Jensen, M.C. and Meckling, W.H. 1976. Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics* 3 (4): 305-360. Doi:10.1016/0304-405x(76)90026-x.

Jeong, I.J., 2011. A dynamic model for the optimization of decoupling point and production planning in a supply chain. *International Journal of Production Economics*, 131(2), pp.561–567. Available at: http://dx.doi.org/10.1016/j.ijpe.2011.02.001.

Jeunet, J., 2006. Demand forecast accuracy and performance of inventory policies under multi-level rolling schedule environments. *International Journal of Production Economics*, 103, pp.401–419.

Ji, G., Gunasekaran, A. and Yang, G., 2014. Constructing sustainable supply chain under double environmental medium regulations. *International Journal of Production Economics*, 147(PART B), pp.211–219. Available at: http://dx.doi.org/10.1016/j.ijpe.2013.04.012.

Jin, Y., Williams, B.D., Tokar, T. and Waller, M.A., 2015. Forecasting With Temporally Aggregated Demand Signals in a Retail Supply Chain. *Journal of Business Logistics*, 36(2), pp.199–211.

Johnson, E.M. and Anderson, E., 2000. Postponement Strategies for Channel Derivatives. *The International Journal of Logistics Management*, 11(1), pp.19–36. Available at: http://www.emeraldinsight.com/doi/10.1108/09574090010806047.

Johnston, F.R. and Harrison, P.J., 1980. An application of forecasting in the alcoholic drinks industry. *Journal of the Operational Research Society*, 31, pp.699–709.

Jöreskog, K.G., 2002. Censored Variables and Censored Regression. *Retrieved January 19*, P.2007.

Juselius, K., 1985. Modelling short- and long-term effects in the aggregate demand for soft drinks. *International Journal of Forecasting*, 1, pp.253–272.

Kahn, K.B., 2003. How to measure the impact of a forecast error on an enterprise. *Journal of Business Forecasting Methods & Systems*, Spring, pp.21–25.

Kannan, G., Grigore, M.C., Devika, K. and Senthilkumar, A., 2013. An analysis of the general benefits of a 216entralized VMI system based on the EOQ model. *International Journal of Production Research*, 51(1), pp.172–188. Available at: http://www.tandfonline.com/doi/abs/10.1080/00207543.2011.653838.

Karl T.R. and Koss W. J., 1984: Historical Climatology Series 4-3: Regional and National Monthly, Seasonal and Annual Temperature Weighted by Area, 1895-1983

Katsikopoulos, K. and Syntetos, A.A., 2016. Bias-Variance Trade-offs in Demand Forecasting. *Foresight: The International Journal of Applied Forecasting*, (40), pp.12–20.

Katz, R.W. and Lazo, J.K., 2011. Economic Value of Weather and Climate Forecasts. In *Oxford Handbooks Online*. pp. 1–32. Available at: http://books.google.com/books?hl=en&lr=&id=FnTVdEfsY2oC&pgis=1.

Katz, R.W. and Murphy, A.H., 1990. Quality/value relationships for imperfect weather forecasts in a prototype multistage decision-making model. *Journal of Forecasting*, 9(1), pp.75–86.

Katz, R.W. and Murphy, A.H. eds., 1997. *Economic Value of Weather and Climate Forecasts*. Cambridge University Press.

Kaufmann, P.J. and Lafontaine, F., 1994. Costs of Control: The Source of Economic Rents for McDonald's Franchisees. *The Journal of Law and Economics* 37 (2): 417-453. doi:10.1086/467319.

Kelle, P. and Silver, E.A., 1989. Forecasting the Returns of Reusable Containers. *Journal of Operations Management*, 8(1), pp.17–35.

Kerkkänen, A., Korpela, J. and Huiskonen, J., 2009. Demand forecasting errors in industrial context: Measurement and impacts. *International Journal of Production Economics*, 118(1), pp.43–48.

Khouja, M., 1998. An aggregate production planning framework for the evaluation of volume flexibility. *Production Planning & Control*, 9(2), pp.127-137.

Khouja, M. and Kumar, R.L., 2002. Information technology investments and volume-flexibility in production systems. *International Journal of Production Research*, 40(1), pp.205–221. Available at: http://www.tandfonline.com/doi/abs/10.1080/00207540110072948.

Kiely, D., 2004. The State of Pharmaceutical Industry Supply Planning and Demand Forecasting. *Journal of Business Forecasting Methods & Systems*, 23(3), pp.20–22. Available at:

http://search.proquest.com/docview/226914243?accountid=10297%5Cnhttp://metalib.dmz.cranfield.ac.uk:9003/cranfield?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=unknown&sid=ProQ:ProQ:abiglobal&atitle=THE+STATE+OF+PHARMACEUTICAL+INDUSTR.

Kim, J.S., Shin, K.Y. and Ahn, S.E., 2003. A Multiple Replenishment Contract with ARIMA Demand Processes. *Journal of the Operational Research Society*, 54(11), pp.1189–1197.

King, M.A., Abrahams, A.S. and Ragsdale, C.T., 2014. Ensemble methods for advanced skier days prediction. *Expert Systems with Applications*, 41, pp.1176–1188.

Kitchin, R. and McArdle, G., 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1), pp.1–10. Available at: http://bds.sagepub.com/lookup/doi/10.1177/2053951716631130.

Koksalan, M., Erkip, N. and Moskowitz, H., 1999. Explaining beer demand: A residual modeling regression approach using statistical process control. *International Journal of Production Economics*, 58(3), pp.265–276. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0925527398002072.

Kolassa, S., 2016. Sometimes it's better to Be Simple than Correct. *Foresight: International Journal of Applied Forecasting*, Winter, pp.20–27.

Kolassa, S., 2016. Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting*, 32(3), pp.788–803. Available at: http://dx.doi.org/10.1016/j.ijforecast.2015.12.004.

Kolassa, S. and Martin, R., 2011. Percentage errors can ruin your day (and rolling the dice shows how). *Foresight: International Journal of Applied Forecasting*, Fall, pp.21–27.

Krueger, A.B., 1991. Ownership, Agency, and Wages: An Examination of Franchising In the Fast Food Industry. *The Quarterly Journal of Economics* 106 (1): 75-101. doi:10.2307/2937907.

Kuhn, M. and Johnson K. 2013. *Applied Predictive Modeling*. Springer. doi:10.1007/978-1-4614-6849-3.

Kung, L.C. and Chen, Y.J., 2014. Impact of Reseller's and Sales Agent's forecasting Accuracy in a Multilayer Supply Chain. *Naval Research Logistics*, 61(3), pp.207–222. Available at: http://www.interscience.wiley.com/jpages/0894-069X/.

Kurtuluş, M., Ülkü, S. and Toktay, B.L., 2012. The Value of Collaborative Forecasting in Supply Chains. *Manufacturing & Service Operations Management*, 14(1), pp.82–98. Available at: http://pubsonline.informs.org/doi/abs/10.1287/msom.1110.0351.

Lackes, R., Schlüter, P. and Siepermann, M., 2016. The impact of contract parameters on the supply chain performance under different power constellations. *International Journal of Production Research*, 54(1), pp.251–264. Available at: http://www.tandfonline.com/doi/full/10.1080/00207543.2015.1076943.

Lafontaine, F., 1992. Agency Theory and Franchising: Some Empirical Results. *The RAND Journal of Economics* 23 (2): 263. doi:10.2307/2555988.

Lazo, J.K., Lawson, M., Larsen, P.H. and Waldman, D.M., 2011. U.S. economic sensitivity to weather variability. *Bulletin of the American Meteorological Society*, 92(6), pp.709–720.

LeBlanc, L., Hill, J. and Harder, J., 2009. Modeling uncertain forecast accuracy in supply chains with postponement. *Journal of Business Logistics*, 30(1), pp.19–31. Available at: http://onlinelibrary.wiley.com/doi/10.1002/j.2158-1592.2009.tb00097.x/full.

Lee, H.L., 1996. Effective inventory and service management through product and process redesign. *Operations Research*, 44(1), pp.151–159.

Lee, H.L., Padmanabhan, V. and Whang, S., 1997. Information distortion in a supply chain: The bullwhip effect. *Management science*, 43(4), pp.546-558.

Lee, H.L., So, K.C. and Tang, C.S., 2000. The Value of Information Sharing in a Two-Level Supply Chain. *Management Science*, 46(5), pp.626–643. Available at: http://www.jstor.org/stable/2661463.

Lee, T.S. and Adam Jr., E.E., 1986. Forecasting Error Evaluation in Material Requirements Planning (MRP) Production-Inventory Systems. *Management Science*, 32(9), pp.1186–1205.

Lee, T.S., Adam Jr., E.E. and Ebert, R.J., 1987. An Evaluation of Forecast Error In Master Production Scheduling for Material Requirements Planning Systems. *Decision Sciences*, 18, pp.292–307.

Lee, T.S., Cooper, F.W. and Adam Jr., E.E., 1993. The Effects of Forecasting Errors on the Total Cost of Operations. *Omega: International Journal of Management Science*, 21(5), pp.541–550.

Li, C., Luo, X., Zhang, C. and Wang, X., 2017. Sunny, Rainy, and Cloudy with a Chance of Mobile Promotion Effectiveness. Marketing Science.

Li, Y., Ye, F. and Lin, Q., 2015. Optimal lead time policy for short life cycle products under Conditional Value-at-Risk criterion. *Computers and Industrial Engineering*, 88, pp.354–365. Available at: http://dx.doi.org/10.1016/j.cie.2015.07.011.

Liao, W.T. and Chang, P.C., 2010. Impacts of forecast, inventory policy, and lead time on supply chain inventory numerical study. *International Journal of Production Economics*, 128(2), pp.527–537. Available at: http://dx.doi.org/10.1016/j.ijpe.2010.07.002.

Lin, N.P., Krajewski, L., Leong, G.K. and Benton, W.C., 1994. The effects of environmental factors on the design of master production scheduling systems. *Journal of Operations Management*, 11(4), pp.367–384.

Lorentz, H., Wong, C.Y. and Hilmola, O.P., 2007. Emerging distribution systems in central and Eastern Europe Implications from two case studies. *International Journal of Physical Distribution & Logistics Management*, 37(8), pp.670–697.

Ma, Y., Wang, N., Che, A., Huang, Y. and Xu, J., 2013. The bullwhip effect under different information-sharing settings: a perspective on price-sensitive demand that incorporates price dynamics. *International Journal of Production Research*, 51(10), pp.3085–3116.

Ma, Y., Wang, N., Che, A., Huang, Y. and Xu, J., 2013. The bullwhip effect on product orders and inventory: a perspective of demand forecasting techniques. *International Journal of Production Research*, 51(1), pp.281–302. Available at: http://www.tandfonline.com/doi/abs/10.1080/00207543.2012.676682.

Mahoney, J.T., 2005. *Economic Foundations of Strategy*. Thousand Oaks, CA: Sage.

Makridakis, S.G., 1993. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9, pp.527–529.

Makridakis, S.G. and Hibon, M., 2000. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16, pp.451–476.

Makridakis, S.G. and Wheelwright, S.C., 1989. *Forecasting methods for management*.

Marmier, F. and Cheikhrouhou, N., 2010. Structuring and integrating human knowledge in demand forecasting: a judgmental adjustment approach. *Production Planning & Control*, 21(4), pp.399–412. Available at: http://www.tandfonline.com/doi/abs/10.1080/09537280903454149.

Masuchun, W., Davis, S. and Patterson, J.W., 2004. Comparison of push and pull control strategies for supply network management in a make-to-stock environment. *International Journal of Production Research*, 42(20), pp.4401–4419.

Mathews, B.P. and Diamantopoulos, A., 1994. Towards a taxonomy of forecast error measures a factor-comparative investigation of forecast error dimensions. *Journal of Forecasting*, 13(4), pp.409–416.

Mathewson, G.F., and Winter, R.A., 1985. The Economics of Franchise Contracts. *The Journal of Law and Economics* 28 (3): 503-526. doi:10.1086/467099.

Maydeu-Olivares, A. and Garcia-Forero, C., 2010. Goodness-of-Fit Testing. *International Encyclopedia of Education*, *7*(1), Pp.190-196.

McAfee, A. and Brynjolfsson, E., 2012. Big Data. The management revolution. *Harvard Business Review*, 90(10), pp.61–68. Available at: http://www.buyukverienstitusu.com/s/1870/i/Big_Data_2.pdf.

McCarthy, T.M. and Golicic, S.L., 2002. Implementing collaborative forecasting to improve supply chain performance. *International Journal of Physical Distribution & Logistics Management*, 32(6), pp.431–454. Available at: http://www.emeraldinsight.com/doi/10.1108/09600030210437960.

Megahed, F.M. and Jones-Farmer, A., 2013. A Statistical Process Monitoring Perspective on "Big Data." *Frontiers in Statistical Quality Control*, 11th ed, p.21. Available at: http://www.eng.auburn.edu/users/fmm0002/ISQC2013Paper.pdf.

Mena, C., Terry, L.A., Williams, A. and Ellram, L., 2014. Causes of waste across multi-tier supply networks: Cases in the UK food sector. *International Journal of Production Economics*, 152, pp.144–158. Available at: http://dx.doi.org/10.1016/j.ijpe.2014.03.012.

Mentzer, J.T. and Cox Jr., J.E., 1984. A Model of the Determinants of Achieved Forecast Accuracy. *Journal of Business Logistics*, 5(2), pp.143–155.

Metters, R., 1997. Quantifying the bullwhip effect in supply chains. *Journal of operations management*, 15(2), pp.89-100.

Michael, S.C., 2000. The Effect of Organizational Form on Quality: The Case of Franchising. *Journal of Economic Behavior & Organization* 43 (3): 295-318. doi:10.1016/s0167-2681(00)00125-6.

Minka, T.P., 2003. A Comparison of Numerical Optimizers for Logistic Regression. *Unpublished Draft*.

Miyaoka, J. and Hausman, W., 2004. How a Base Stock Policy Using "Stale" Forecasts Provides Supply Chain Benefits. *Manufacturing & Service Operations Management*, 6(2), pp.149–162.

Miyaoka, J. and Hausman, W.H., 2008. How Improved Forecasts Can Degrade Decentralized Supply Chains. *Manufacturing & Service Operations Management*, 10(3), pp.547–562.

Mjelde, J.W., Sonka, S.T. and Peel, D.S., 1989. Climate and Weather forecasting: A Review. *Water*, 7495(August), p.22.

Mjelde, J.W. and Dixon, B.L., 1993. Valuing the lead time of periodic forecasts in dynamic production systems. Agricultural Systems, 42(1–2), pp.41–55.

Moon, M.A., 2015. Commentary on Improving Forecast Quality in Practice. *Foresight: The International Journal of Applied Forecasting*, 25(1), p.24. Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001577-201325010-00006.

Morlidge, S., 2014. Forecast Quality in the Supply Chain. *Foresight: International Journal of Applied Forecasting*, Spring, pp.26–32.

Morlidge, S., 2013. How Good Is a "Good" Forecast? Forecast Errors and Their Avoidability. *Foresight: International Journal of Applied Forecasting*, Summer, pp.5–12.

Morlidge, S., 2014. Using Relative Error Metrics to Improve Forecast Quality in the Supply Chain. *Foresight: The International Journal of Applied Forecasting*, Summer (34), pp.39–47.

Murphy, A.H., 1977. The Value of Climatological, Categorical and Probabilistic Forecasts in the Cost-Loss Ratio Situation. *Monthly Weather Review*, 105(7), pp.803–816.

Murphy, A.H., 1993. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting*, 8(2), pp.281–293.

Murphy, A.H., 1997. Forecast Verification. *Economic value of weather and climate forecasts*, pp.19-74.

Murray, K.B., Di Muro, F., Finn, A. and Leszczyc, P.P., 2010. The effect of weather on consumer spending. *Journal of Retailing and Consumer Services*, 17(6), pp.512–520. Available at: http://dx.doi.org/10.1016/j.jretconser.2010.08.006.

Nagashima, M., Wehrle, F.T., Kerbache, L. and Lassagne, M., 2015. Impacts of adaptive collaboration on demand forecasting accuracy of different product categories throughout the product life cycle. *Supply Chain Management: An International Journal*, 20(4), pp.415–433.

National Research Council, 2013. Frontiers in Massive Data Analysis. Technical Paper No. 201402. *Board on Mathematical Sciences and Their Applications, National Research Council*. Washington, D.C.: National Academies Press.

NCEI (National Center for Environmental Information)., 2017. U.S. Local Climatological Data. Rep. Asheville, NC. Web. https://www.ncdc.noaa.gov/orders/qclcd/.

NGIA (National Geospatial Intelligence Agency)., 2014. World Geodetic System 1984 (WGS 84). Office of Geomatics. http://earth-info.nga.mil/GandG/wgs84/index.html.

Niakan, F. and Rahimi, M., 2015. A multi-objective healthcare inventory routing problem; a fuzzy possibilistic approach. *Transportation Research Part E: Logistics and Transportation Review*, 80, pp.74–94. Available at: http://dx.doi.org/10.1016/j.tre.2015.04.010.

Nikolopoulos, K. and Fildes, R., 2013. Adjusting supply chain forecasts for short-term temperature estimates: A case study in a Brewing company. *IMA Journal of Management Mathematics*, 24(1), 79–88. https://doi.org/10.1093/imaman/dps006.

Noren, D.L., 1990. The Economics Of The Golden Arches: A Case Study of the McDonald's System. *The American Economist* 34 (2 (Fall 1990): 60-64.

Norton, S.W., 1988. Franchising, Brand Name Capital, And The Entrepreneurial Capacity Problem. *Strategic Management Journal*. 9 (S1): 105-114. doi:10.1002/smj.4250090711.

Norton, S.W., 1988. An Empirical Look at Franchising As an Organizational Form. *The Journal of Business* 61 (2): 197. doi:10.1086/296428.

Oliva, R. and Watson, N., 2009. Managing Functional Biases in Organizational Forecasts: A Case Study of Consensus Forecasting in Supply Chain Planning. *Production and Operations Management*, 18(2), pp.138–151.

OSC., 1987. Ohio Supercomputer Center. Columbus OH. http://osc.edu/ark:/19495/f5s1ph73.

Oxenfeldt, A.R. and Kelly, A.O., 1969. Will Successful Franchise Systems Ultimately Become Wholly-Owned Chains? *Journal Of Retailing* 44 (4): 69-83.

Paik, Y. and Choi, D.Y., 2007. Control, Autonomy and Collaboration in the Fast Food Industry: A Comparative Study between Domestic and International Franchising. *International Small Business Journal* 25 (5): 539-562. doi:10.1177/0266242607080658.

Palmer, B., 2013. Long-Term Weather Forecasts Are A Long Way From Accurate. *Washington Post*. https://www.washingtonpost.com/national/health-science/long-term-

weather-forecasts-are-a-long-way-from-accurate/2013/04/15/1f9a2ac8-a05b-11e2-be47-b44febada3a8_story.html?utm_term=.8434156533de.

Parsons, A.G., 2001. The Association between Daily Weather and Daily Shopping Patterns. *Australasian Marketing Journal* (AMJ), 9(2), pp.78–84. Available at: http://dx.doi.org/10.1016/S1441-3582(01)70177-2.

Paul, A., Tan, Y. and Vakharia, A.J., 2015. Inventory Planning for a Modular Product Family. *Production and Operations Management*, 24(7), pp.1033–1053.

Pearson, M., Masson, R. and Swain, A., 2010. Process control in an agile supply chain network. *International Journal of Production Economics*, 128(1), pp.22–30. Available at: http://dx.doi.org/10.1016/j.ijpe.2010.01.027.

Penrose, E.T., 1959. *The Theory of the Growth of the Firm*. Oxford: Oxford University Press.

Persinger, M. and Levesque, B.F., 1983. Geophysical variables and behavior: XII. The weather matrix accommodates large portions of variance of measured daily mood. *Perceptual and Motor Skills*, 57, pp.868–870.

Persona, A., Battini, D., Manzini, R. and Pareschi, A., 2007. Optimal safety stock levels of subassemblies and manufacturing components. *International Journal of Production Economics*, 110(1–2), pp.147–159.

Peter, Ď. and Silvia, P., 2012. ARIMA vs. ARIMAX – which approach is better to analyze and forecast macroeconomic time series? In *Proceedings of 30th International Conference Mathematical Methods in Economics θ*. pp. 136–140.

Petersen, H., 2003. Integrating the Forecasting Process with the Supply Chain: Bayer Healthcare's Journey. *Journal of Business Forecasting Methods & Systems*, 22(4), pp.11–16.

Potgieter, A.B., Everingham, Y.L. and Hammer, G.L., 2003. On measuring quality of a probabilistic commodity forecast for a system that incorporates seasonal climate forecasts. *International Journal of Climatology*, 23(10), pp.1195–1210.

Rajaram, K. and Tang, C.S., 2001. The impact of product substitution on retail merchandising. *European Journal of Operational Research*, 135(3), pp.582-601.

Rajesh, R. and Ravi, V., 2015. Modeling enablers of supply chain risk mitigation in electronic supply chains: A Grey-DEMATEL approach. *Computers and Industrial Engineering*, 87, pp.126–139. Available at: http://dx.doi.org/10.1016/j.cie.2015.04.028.

Ramanathan, U., 2013. Aligning supply chain collaboration using Analytic Hierarchy Process. *Omega: International Journal of Management Science*, 41(2), pp.431–440. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0305048312000552.

Ramanathan, U. and Muyldermans, L., 2010. Identifying demand factors for promotional planning and forecasting: A case of a soft drink company in the UK. *International Journal of Production Economics*, 128(2), pp.538–545. Available at: http://dx.doi.org/10.1016/j.ijpe.2010.07.007.

Ritzman, L.P. and King, B.E., 1993. The relative significance of forecast errors in multistage manufacturing. *Journal of Operations Management*, 11(1), pp.51–65.

Ritzman, L.P. and Sanders, N.R., 2001. Judgmental Adjustment of Statistical Forecasts. In J. S. Armstrong, ed. *Principles of Forecasting*. New York: Springer Science+ Business Media.

Rodrigues, P.M.M. and Osborn, D.R., 1999. Performance of seasonal unit root tests for monthly data. *Journal of Applied Statistics*, 26(8), p.985± 1004.

Rostami-Tabar, B., Babai, M.Z., Syntetos, A.A. and Ducq, Y., 2013. Demand Forecasting by Temporal Aggregation. *Naval Research Logistics*, 60(6), pp.479–498. Available at: http://www.interscience.wiley.com/jpages/0894-069X/.

Rubin, P.H., 1978. The Theory of the Firm and the Structure of the Franchise Contract. *The Journal of Law and Economics* 21 (1): 223-233. doi:10.1086/466918.

Samenow, J. and Fritz, A., 2015. Five Myths about Weather Forecasting. *Washington Post*. Available at: https://www.washingtonpost.com/opinions/five-myths-about-weather-forecasting/2015/01/02/e49e8950-8b86-11e4-a085-34e9b9f09a58_story.html?utm_term=.18b10a1622a0.

Sanchez-Lugo, E., 2017. U.S. Climate Regions. *National Climatic Data Center, NOAA*, www.ncdc.noaa.gov/monitoring-references/maps/us-climate-regions.php.

Sanders, J.L. and Brizzolara, M.S., 1982. Relationships between weather and mood. *The Journal of General Psychology*, 107(1), pp.155-156.

Sanders, N.R. and Graman, G.A., 2009. Quantifying costs of forecast errors: A case study of the warehouse environment. *Omega: International Journal of Management Science*, 37(1), pp.116–125.

Sanders, N.R. and Graman, G.A., 2016. Impact of Bias Magnification on Supply Chain Costs: The Mitigating Role of Forecast Sharing. *Decision Sciences*, 47(5), pp.881–906.

Sanders, N.R. and Manrodt, K.B., 2003. The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega: International Journal of Management Science*, 31(6), pp.511–522.

Sanders, N.R. and Ritzman, L.P., 2004. Using Warehouse Workforce Flexibility to Offset Forecast Errors. *Journal of Business Logistics*, 25(2), pp.251–270.

Sanders, N.R. and Ritzman, L.P., 2004. Integrating judgmental and quantitative forecasts: methodologies for pooling marketing and operations information. *International Journal of Operations & Production Management*, 24(5), pp.514-529.

Sanders, N.R. and Ritzman, L.P., 1995. Bringing judgment into combination forecasts. *Journal of Operations Management*, 13(4), pp.311–321.

Schmitt, T.G., 1984. Resolving uncertainty in manufacturing systems. *Journal of Operations Management*, 4(4), pp.331–345.

Sethi, S.P., Yan, H., Zhang, H. and Zhou, J., 2007. A supply chain with a service requirement for each market signal. *Production and Operations Management*, 16(3), pp.322–342.

Sharma, A. and Paliwal, K.K., 2007. Fast Principal Component Analysis Using Fixed-Point Algorithm. *Pattern Recognition Letters*, 28(10), Pp.1151-1155.

Shelton, J.P., 1967. Allocative Efficiency Vs." X-Efficiency": Comment. *The American Economic Review* 57 (5): 1252-1258.

Shin, H. and Tunca, T.I., 2010. Do Firms Invest in Forecasting Efficiently? The Effect of Competition on Demand Forecast Investments and Supply Chain Coordination. *Operations Research*, 58(6), pp.1592–1610.

Silver, E.A., Pyke, D.F. and Peterson, R., 1998. *Inventory Management and Production Planning and Scheduling* (Vol. 3, p. 30). New York: Wiley.

Silver, N. 2012. *The signal and the noise: Why so many predictions fail-but some don't.* Penguin.

Smith, C.D. and Mentzer, J.T., 2010. User Influence on the Relationship between Forecast Accuracy, Application and Logistics Performance. *Journal of Business Logistics*, 31(1), pp.159–177.

Smith, K., 1993. The influence of weather and climate on recreation and tourism. *Weather*, 48(12), pp.398–404.

Smyth, K.B., Croxton, K.L., Franklin, R. and Knemeyer, A.M., 2017. Thirsty in an Ocean of Data? Pitfalls and Practical Strategies when Partnering with Industry on Big

Data Supply Chain Research. *Journal of Business Logistics*. Manuscript submitted for publication (copy on file with author).

Sourirajan, K., Ramachandran, B. and An, L., 2008. Application of control theoretic principles to manage inventory replenishment in a supply chain. *International Journal of Production Research*, 46(21), pp.6163–6188. Available at: http://www.tandfonline.com/doi/abs/10.1080/00207540601178151.

Southern, R.N., 2011. Historical perspective of the logistics and supply chain management discipline. *Transportation Journal*, 50(1), pp.53-64.

Spies, K., Hesse, F. and Loesch, K., 1997. Store atmosphere, mood and purchasing behavior. *International Journal of Research in Marketing*, 14, pp.1–17.

Sridharan, V. and LaForge, L.R., 1989. The impact of safety stock on schedule instability, cost and service. *Journal of Operations Management*, 8(4), pp.327–347.

Starr-McCluer, M., 2000. The Effects of Weather on Retail Sales. Finance and Economics Discussion Series *Working Paper*, 2000–8(Federal Reserve Board of Governors).

Steele, A.T., 1951. Weather's Effect on the Sales of a Department Store. *Journal of Marketing*, 15(4), pp.436–443.

Steinker, S., Hoberg, K. and Thonemann, U.W., 2016. The Value of Weather Information for E-commerce Operations. *Production and Operations Management*.

Stern, H. and Davidson, N.E., 2015. Trends in the skill of weather prediction at lead times of 1 – 14 days. *Quarterly Journal of the Royal Meteorological Society*, (October), pp.2726–2736.

Stockton, N., 2016. Deep Thunder can forecast the weather – down to a city block. *Wired Magazine*. https://www.wired.com/2016/06/deep-thunder-can-forecast-weather-city-block/.

Suddath, C., 2014. The Weather Channel's Secret: Less Weather, More Clickbait. *Bloomberg*. https://www.bloomberg.com/news/articles/2014-10-09/weather-channels-web-mobile-growth-leads-to-advertising-insights.

Syntetos, A.A., Babai, Z., Boylan, J.E., Kolassa, S. and Nikolopoulos, K., 2016. Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, 252(1), pp.1-26.

Syntetos, A.A. and Boylan, J.E., 2001. On the bias of intermittent demand estimates. *International Journal of Production Economics*, 71, pp.457–466.

Syntetos, A.A. and Boylan, J.E., 2005. The accuracy of intermittent demand estimates. *International Journal of Forecasting*, 21, pp.303–314.

Syntetos, A.A. and Boylan, J.E., 2010. On the variance of intermittent demand estimates. *International Journal of Production Economics*, 128(2), pp.546–555. Available at: http://dx.doi.org/10.1016/j.ijpe.2010.07.005.

Syntetos, A.A., Boylan, J.E. and Croston, J.D., 2005. On the categorization of demand patterns. *Journal of the Operational Research Society*, 56(5), pp.495-503.

Syntetos, A.A., Boylan, J.E. and Disney, S.M., 2009. Forecasting For Inventory Planning: A 50-Year Review. *Journal of the Operational Research Society.* 60: S149-S160. doi:10.1057/jors.2008.173.

Syntetos, A.A., Nikolopoulos, K. and Boylan, J.E., 2010. Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *International Journal of Forecasting*, 26(1), pp.134–143. Available at: http://dx.doi.org/10.1016/j.ijforecast.2009.05.016.

Tate, W.L., Ellram, L.M. and Dooley, K.J., 2012. Environmental purchasing and supplier management (EPSM): Theory and practice. *Journal of Purchasing and Supply Management*, 18(3), pp.173-188.

Taylor, T.A. and Xiao, W., 2010. Does a Manufacturer Benefit from Selling to a Better-Forecasting Retailer? *Management Science*, 56(9), pp.1584–1598.

Terwiesch, C., Ren, Z.J., Ho, T.H. and Cohen, M.A., 2005. An Empirical Analysis of Forecast Sharing in the Semiconductor Equipment Supply Chain. *Management Science*, 51(2), pp.208–220. Available at: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1040.0317.

Thiel, D., Vo, T.L.H. and Hovelaque, V., 2014. Forecasts impacts on sanitary risk during a crisis: a case study. *The International Journal of Logistics Management*, 25(2), pp.358–378.

Theil, H., 1966. *Applied economic forecasting*. Chicago, IL7 Rand McNally.

Thompson, J.C. and Brier, G.W., 1955. The Economic Utility of Weather Forecasts. *Monthly Weather Review*, 83(11), pp.249–254.

Thompson, J.C., 1962. Economic Gains from Scientific Advances and Operational Improvements in Meteorological Prediction. *Journal of Applied Meteorology*, 1(March).

Tibben-lembke, R.S. and Amato, H.N., 2001. Replacement Parts Management: The Value of Information. *Journal of Business Logistics*, 22(2), pp.149–165.

Toulis, P. and Airoldi, E.M., 2015. Scalable Estimation Strategies Based On Stochastic Approximations: Classical Results and New Insights. *Statistics and Computing*, *25*(4), Pp.781-795.

Tran, B. R., 2016. Blame it on the Rain - Weather Shocks and Retail Sales. Retrieved from http://econweb.ucsd.edu/~brothtra/pdfs/BlameItOnTheRain.pdf.

Tranfield, D., Denyer, D. and Smart, P., 2003. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, 14(3), pp.207-222.

Trapero, J.R., Kourentzes, N. and Fildes, R., 2012. Impact of information exchange on supplier forecasting performance. *Omega: International Journal of Management Science*, 40(6), pp.738–747. Available at: http://dx.doi.org/10.1016/j.omega.2011.08.009.

Tratar, L.F., 2010. Joint optimisation of demand forecasting and stock control parameters. *International Journal of Production Economics*, 127(1), pp.173–179. Available at: http://dx.doi.org/10.1016/j.ijpe.2010.05.009.

Tribbia, J.J., 1997. Weather prediction. *Economic value of weather and climate forecasts*, pp.1-12.

Tucker, P. and Gilliland, J., 2007. The effect of season and weather on physical activity: A systematic review. *Public Health*, 121, pp.909–922.

Ülkü, S., Toktay, L.B. and Yücesan, E., 2007. Risk Ownership in Contract Manufacturing. *Manufacturing & Service Operations Management*, 9(3), pp.225–241. Available at: http://pubsonline.informs.org/doi/abs/10.1287/msom.1060.0141.

van der Laan, E., van Dalen, J., Rohrmoser, M. and Simpson, R., 2016. Demand forecasting and order planning for humanitarian logistics: An empirical assessment. *Journal of Operations Management*, 45, pp.114–122. Available at: http://dx.doi.org/10.1016/j.jom.2016.05.004.

Velicer, W.F., 1976. Determining the Number of Components from the Matrix of Partial Correlations. *Psychometrika*, 41(3), Pp.321-327.

Venkataraman, R. and D'Itri, M.P., 2001. Rolling horizon master production schedule performance: a policy analysis. *Production planning & control*, 12(7), pp.669-679.

Venkataraman, R. and Nathan, J., 1999. Effect of forecast errors on rolling horizon master production schedule cost performance for various replanning intervals. *Production Planning & Control*, 10(7), pp.682-689.

Vose, R.S., Applequist, S., Durre, I., Menne, M.J., Williams, C.N., Fenimore, C., Gleason, K. and Arndt, D., 2014: Improved Historical Temperature and Precipitation Time Series For U.S. Climate Divisions. *Journal of Applied Meteorology and Climatology*. DOI: http://dx.doi.org/10.1175/JAMC-D-13-0248.1

Wacker, J.G. and Sprague, L.G., 1995. The impact of institutional factors on forecast accuracy: manufacturing executives' perspective. *International Journal of Production Research*, 33(11), pp.2945–2958.

Wacker, J.G. and Sprague, L.G., 1998. Forecasting accuracy: comparing the relative effectiveness of practices between seven developed countries. *Journal of Operations Management*, 16, pp.271–290.

Waller, M.A. and Fawcett, S.E., 2013. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2), pp.77–84.

Waller, M.A. and Fawcett, S.E., 2013. Click Here for a Data Scientist: Big Data, Predictive Analytics, and Theory Development in the Era of a Maker Movement Supply Chain. *Journal of Business Logistics*, 34(4), pp.249–252.

Wallstrom, P. and Segerstedt, A., 2010. Evaluation of forecasting error measurements and techniques for intermittent demand. *International Journal of Production Economics*, 128, pp.625–636.

Wan, X. and Sanders, N.R., 2017. The negative impact of product variety: Forecast bias, inventory levels, and the role of vertical integration. *International Journal of Production Economics*, 186(July 2016), pp.123–131. Available at: http://dx.doi.org/10.1016/j.ijpe.2017.02.002.

Wemmerlöv, U., 1985. Comments on "Cost Increases Due To Demand Uncertainty in MRP Lot Sizing": A Verification of Ordering Probabilities. *Decision Sciences*, 16(4), pp.410–419.

Wemmerlöv, U., 1989. The behavior of lot-sizing procedures in the presence of forecast errors. *Journal of Operations Management*, 8(1), pp.37–47.

Wemmerlöv, U., 1984. Assemble-To-Order Manufacturing: Implications for Materials Management. *Journal of Operations Management*, 4(4), pp.347–368.

Werdigier, J., 2009. British Grocery Chain Uses The Weather to Predict Sales. *New York Times*, September 1, p. B6.

Wickramatillake, C.D., Koh, L.S.C., Gunasekaran, A. and Arunachalam, S., 2007. Measuring performance within the supply chain of a large scale project. *Supply Chain*

Management: An International Journal, 12(1), pp.52–59. Available at: http://www.emeraldinsight.com/doi/10.1108/13598540710724338.

Widiarta, H., Viswanathan, S. and Piplani, R., 2006. On the Effectiveness of Top-Down Strategy for Forecasting Autoregressive Demands. *Naval Research Logistics*, 54(Mar 2007), pp.176–188. Available at: http://www.interscience.wiley.com/jpages/0894-069X/.

Wieland, A. and Wallenburg, C.M., 2012. Dealing with supply chain risks: Linking risk management practices and strategies to performance. *International Journal of Physical Distribution & Logistics Management*, 42(10), pp.887-905.

Willemain, T.R., Smart, C.N. and Schwarz, H.F., 2004. A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting*, 20(3), pp.375–387.

Williams, B.D. and Tokar, T., 2008. A review of inventory management research in major logistics journals: Themes and future directions. *The International Journal of Logistics Management*, 19(2), pp.212-232.

Williams, B.D. and Waller, M.A., 2011. Estimating a retailer's base stock level: an optimal distribution center order forecast policy. *Journal of the Operational Research Society*, 62(4), pp.662–666.

Williams, B.D. and Waller, M.A., 2011. Top-down versus bottom-up demand forecasts: The value of shared point-of-sale data in the retail Supply Chain. *Journal of Business Logistics*, 32(1), pp.17–26.

Williams, B.D. and Waller, M.A., 2010. Creating Order Forecasts: Point-of-Sale or Order History? *Journal of Business Logistics*, 31(2), pp.231–251. Available at: http://doi.wiley.com/10.1002/j.2158-1592.2010.tb00150.x.

Winklhofer, H., Diamantopoulos, A. and Witt, S., 1996. Forecasting practice: A review of the empirical literature and an agenda for future research. *International Journal of Forecasting*, 12(2), pp.193–221. Available at: http://www.sciencedirect.com/science/article/pii/0169207095006478.

Winter, M. and Knemeyer, A.M., 2013. Exploring the integration of sustainability and supply chain management: Current state and opportunities for future inquiry. *International Journal of Physical Distribution & Logistics Management*, 43(1), pp.18-38.

Wooldridge, J.M., 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wooldridge, J.M., 2015. *Introductory Econometrics: A Modern Approach*. Nelson Education.

Xie, J., Lee, T.S. and Zhao, X., 2004. Impact of forecasting error on the performance of capacitated multi-item production systems. *Computers & Industrial Engineering*, 46, pp.205–219.

Yang, K.K. and Jacobs, F.R., 1999. Replanning the Master Production Schedule for a Capacity-Constrained Job Shop. *Decision Sciences*, 30(3), pp.719–748.

Yao, Y., Kohli, R., Sherer, S.A. and Cederlund, J., 2013. Learning curves in collaborative planning, forecasting, and replenishment (CPFR) information systems: An empirical analysis from a mobile phone manufacturer. *Journal of Operations Management*, 31(6), pp.285–297.

Yao, D.Q., Yue, X., Wang, X. and Liu, J.J., 2005. The impact of information sharing on a returns policy with the addition of a direct channel. *International Journal of Production Economics*, 97(2), pp.196–209.

Yelland, P.M., 2010. Bayesian forecasting of parts demand. *International Journal of Forecasting*, 26(2), pp.374–396. Available at: http://dx.doi.org/10.1016/j.ijforecast.2009.11.001.

Yin, X. and Zajac, E.J., 2004. The Strategy/Governance Structure Fit Relationship: Theory and Evidence in Franchising Arrangements. *Strategic Management Journal.* 25 (4): 365-383. doi:10.1002/smj.389.

Yokum, J.T. and Armstrong, J.S., 1995. Beyond accuracy: Comparison of criteria used to select forecasting methods. *International Journal of Forecasting*, 11(4), pp.591-597.

Zhao, X., Lai, F. and Lee, T.S., 2001. Evaluation of safety stock methods in multilevel material requirements planning (MRP) systems. *Production Planning & Control*, 12(8), pp.794-803.

Zhao, X. and Lee, T.S., 1993. Freezing the master production schedule for material requirements planning systems under demand uncertainty. *Journal of Operations Management*, 11, pp.185–205.

Zhao, X. and Xie, J., 2002. Forecasting errors and the value of information sharing in a supply chain. *International Journal of Production Research*, 40(2), pp.311–335.

Zhao, X., Xie, J. and Leung, J., 2002. The impact of forecasting model selection on the value of information sharing in a supply chain. *European Journal of Operational Research*, 142(2), pp.321–344.

Zhao, X., Xie, J. and Wei, J.C., 2002. The Impact of Forecast Errors on Early Order Commitment in a Supply Chain. *Decision Sciences*, 33(2), pp.251–280. Available at: http://doi.wiley.com/10.1111/j.1540-5915.2002.tb01644.x.

Zhu, X., Mukhopadhyay, S.K. and Yue, X., 2011. Role of forecast effort on supply chain profitability under various information sharing scenarios. *International Journal of Production Economics*, 129(2), pp.284–291. Available at: http://dx.doi.org/10.1016/j.ijpe.2010.10.021.

Zinn, W. and Marmorstein, H. Comparing two alternative methods of determining safety stock levels: The demand and the forecast systems. *Journal of Business Logistics* 11, no. 1 (1990): 95-110.

Zivin, J.G. and Neidell, M., 2014. Temperature and the Allocation of Time: Implications for Climate Change. *Journal of Labor Economics*, 32(1).

Zotteri, G., Kalchschmidt, M. and Caniato, F., 2005. The impact of aggregation level on forecasting performance. *International Journal of Production Economics*, 94, pp.479–491.

Zotteri, G., Kalchschmidt, M. and Saccani, N., 2014. Forecasting by Cross-Sectional Aggregation. *Foresight: The International Journal of Applied Forecasting*, 35, pp.35–42.

Zwick, W.R. and Velicer, W.F., 1986. Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin, 99*(3), P.432.

Appendix A: Boolean Keywords for EBSCO Business Source Complete

The following keywords were applied to Publication Name, Title, Subject, Keywords or Abstract with the EBSCO Business Source Complete databases:

((forecast* N5 (error OR accura* OR quality OR deviation* OR performance OR consistency OR reliability OR precision OR bias))
AND
((supply N5 chain*) OR logistic*)
AND
(((bull* N5 *whip) OR "industrial dynamics" OR "system dynamics")
OR ((trade* N5 *off) OR (break* N5 *even) OR cost* OR "economic impact")
OR (aggregat* OR hierarchical OR combination OR composite OR synthesis OR consensus OR pooling)
OR (judgement* OR subject* OR adjust*)
OR (information AND shar*)
OR (metric* OR measure)
OR (over* N5 *fit*)
OR (((safety OR buffer) N5 (stock OR inventory)) OR "lead time" OR "stock policy" OR inventory)
OR ("supply chain" N5 (collaborat* OR coordinat* OR manage*))
OR ("supply chain" N5 (integrat* OR synchroniz*))
OR ("service level" OR "fill rate" OR "ready rate" OR (stock* AND *out))
OR (horizon OR range)
OR (substitut* OR flexib* OR adapt* OR agil*)
OR (tolerance OR resilience OR robust* OR point)
OR (varia* OR uncertain* OR *predictab* OR *forecastab* OR volatil*)))