Journal of Banking and Financial Dynamics
Vol. 9, No. 12, 1-9, 2025
ISSN(E): 2576-6821
DOI: 10.55220/2576-6821.v9.796

Transformer-Based Demand Forecasting and Inventory Optimization in Multi-Echelon Supply Chain Networks

Xuguang Zhang[1] ✉
Tiejiang Sun[2]
Xu Han[3]
Yongbin Yang[4]
Pan Li[5]

[1]*University of Gloucestershire, Cheltenham, UK.*
[2]*Chang'an University, Xi'an, China.*
[3]*Renmin University of China, China.*
[4]*University of Southern California, Los Angeles, USA.*
[5]*University of Hull, Hull, UK.*
(✉ Corresponding Author)

Abstract

Multi-echelon supply chain networks face increasing complexity in demand forecasting and inventory optimization due to volatile market conditions and dynamic customer preferences. Traditional forecasting methods often struggle to capture long-range dependencies and complex temporal patterns inherent in supply chain data. Transformer-based architectures, originally developed for natural language processing (NLP), have emerged as powerful tools for time series forecasting in supply chain management (SCM). These models leverage self-attention mechanisms to process sequential data and capture intricate relationships across multiple time steps. This review examines the application of transformer models in demand forecasting and inventory optimization within multi-echelon supply chain networks. The paper analyzes how transformer architectures address challenges such as bullwhip effect amplification, demand volatility, and coordination across supply chain tiers. Key findings indicate that transformer-based approaches outperform conventional methods including autoregressive integrated moving average (ARIMA), long short-term memory (LSTM) networks, and traditional machine learning (ML) algorithms in forecast accuracy and computational efficiency. The review synthesizes recent developments in transformer architectures specifically adapted for supply chain contexts, including modifications for handling sparse data, incorporating external factors, and enabling real-time decision support. Furthermore, the paper explores integration strategies between demand forecasting and inventory optimization, examining how transformer predictions inform safety stock calculations, reorder point determination, and dynamic replenishment policies. Emerging trends such as attention mechanism interpretability, federated learning (FL) for collaborative forecasting, and hybrid models combining transformers with reinforcement learning (RL) are discussed. The review identifies critical gaps in current research, including limited validation in real-world multi-echelon settings, computational scalability challenges, and the need for robust frameworks addressing demand uncertainty propagation across supply chain levels.

Keywords: Attention mechanism, Deep learning, Demand forecasting, Inventory optimization, Multi-echelon supply chain, Supply chain management, Time series prediction, Transformer architecture.

1. Introduction

Modern supply chain networks operate in increasingly complex and uncertain environments characterized by fluctuating customer demands, shortened product lifecycles, and heightened competition across global markets. Multi-echelon supply chain networks, consisting of multiple interconnected stages from raw material suppliers to end customers, require sophisticated coordination mechanisms to maintain operational efficiency while minimizing costs and meeting service level requirements [1]. Demand forecasting serves as a critical foundation for effective inventory management, production planning, and resource allocation across these interconnected echelons [2]. However, traditional forecasting approaches often fail to adequately capture the nonlinear dynamics, seasonal variations, and long-term dependencies that characterize modern supply chain demand patterns [3].

The emergence of deep learning (DL) techniques has revolutionized time series forecasting across various domains, offering unprecedented capabilities to model complex temporal relationships and extract meaningful patterns from large-scale historical data [4]. Among these advances, transformer-based architectures have demonstrated remarkable success in natural language processing (NLP) tasks and have recently been adapted for time series forecasting applications [5]. The transformer model employs self-attention mechanisms that enable the model to weigh the importance of different time steps when making predictions, thereby capturing both short-term fluctuations and long-range dependencies without the sequential processing constraints of recurrent neural networks (RNN) [6]. This architectural innovation addresses fundamental limitations of conventional forecasting methods and recurrent architectures in handling long sequences and parallel computation [7].

In the context of multi-echelon supply chain management (SCM), accurate demand forecasting directly influences inventory optimization decisions at each tier of the network [8]. The bullwhip effect, characterized by demand variability amplification as information propagates upstream through supply chain levels, poses significant challenges for inventory planning and can result in excessive safety stocks, stockouts, and increased operational costs [9]. Transformer-based forecasting models offer potential solutions to mitigate these challenges by providing more accurate demand predictions and capturing cross-echelon dependencies through attention mechanisms [10]. This review comprehensively examines the state-of-the-art developments in transformer-based demand forecasting and inventory optimization for multi-echelon supply chains, synthesizes methodological advances, and identifies future research directions.

2. Literature Review

The application of advanced forecasting techniques in supply chain contexts has evolved significantly over the past decades, transitioning from statistical methods to machine learning (ML) and subsequently to DL approaches. Classical time series forecasting methods such as autoregressive integrated moving average (ARIMA) models have been widely employed in supply chain demand prediction due to their mathematical tractability and interpretability [11]. However, these linear models exhibit limited capacity to capture complex nonlinear patterns and multivariate dependencies inherent in modern supply chain data streams [12]. The advent of ML techniques introduced new possibilities for demand forecasting through algorithms such as support vector machines, random forests, and gradient boosting methods, demonstrating improved performance over traditional statistical models by capturing nonlinear relationships [13].

The emergence of DL architectures marked a paradigm shift in time series forecasting capabilities, with RNN and their variants, particularly long short-term memory (LSTM) networks and gated recurrent units (GRU), achieving substantial improvements in capturing temporal dependencies and sequential patterns [14]. These architectures addressed the vanishing gradient problem that limited traditional RNN performance and enabled modeling of longer temporal sequences [15]. Despite their successes, LSTM and GRU architectures face computational inefficiencies due to their inherently sequential nature, which prevents parallel processing of input sequences and limits scalability for large datasets [16]. Furthermore, these recurrent architectures may still struggle with very long-range dependencies spanning hundreds of time steps, which are common in supply chain scenarios involving seasonal patterns and cyclical demand fluctuations [17].

Convolutional neural networks (CNN) have been explored as alternatives for time series forecasting, offering parallel computation capabilities and effective feature extraction through convolutional operations [18]. The transformer architecture revolutionized sequence modeling by introducing self-attention mechanisms that compute relationships between all positions in a sequence simultaneously, enabling both parallel computation and effective capture of long-range dependencies [19]. Original transformer models designed for NLP tasks have been adapted for time series forecasting through various modifications addressing the specific characteristics of temporal data [20]. The Temporal Fusion Transformer introduced by Lim and colleagues specifically targets multi-horizon forecasting problems by incorporating variable selection networks and interpretable attention mechanisms [21]. This architecture demonstrated superior performance across multiple real-world forecasting datasets and provided insights into which input variables and time steps contribute most significantly to predictions [22].

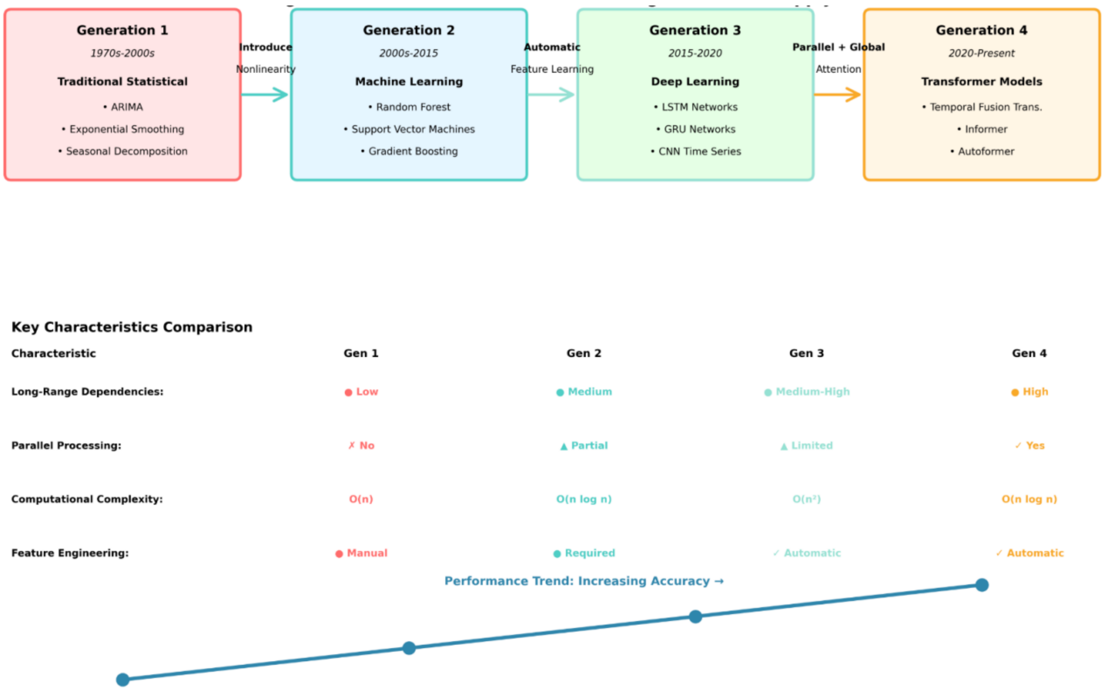
Subsequent research has developed specialized transformer variants for time series applications, including Informer, which addresses computational and memory constraints through ProbSparse self-attention mechanisms and distilling operations that reduce sequence length progressively [23]. The Autoformer architecture incorporates decomposition capabilities that separate trend and seasonal components within the attention mechanism itself, enabling more effective modeling of complex temporal patterns [24]. Pyraformer introduces pyramidal attention structures that capture temporal dependencies at multiple resolutions, reducing computational complexity while maintaining forecasting accuracy [25]. These architectural innovations demonstrate the ongoing evolution of transformer-based approaches specifically designed for time series forecasting challenges.

In supply chain contexts, transformer models have been applied to various forecasting tasks including demand prediction, inventory level forecasting, and supply chain disruption prediction [26]. Recent studies have explored the integration of transformer architectures with domain-specific features such as promotional calendars, holiday effects, and external market indicators to enhance forecasting accuracy for retail and e-commerce supply chains [27]. Multi-task learning frameworks combining transformers with auxiliary prediction objectives have shown promise in improving forecast robustness and capturing relationships between related products or locations [28]. Attention mechanism visualization and interpretability techniques have been employed to understand which historical periods and features most strongly influence demand predictions, providing valuable insights for supply chain planners [29].

The multi-echelon nature of supply chain networks introduces additional complexities for demand forecasting, as demand patterns at different echelons exhibit varying characteristics and dependencies [30]. Upstream echelons typically experience greater demand variability due to the bullwhip effect, necessitating forecasting approaches that account for demand amplification and information distortion across supply chain tiers [31]. Hierarchical forecasting methods that ensure consistency between aggregate and disaggregate predictions have been explored in conjunction with transformer models to maintain coherence across supply chain levels [32]. Graph neural networks (GNN) combined with transformer architectures have been proposed to model supply chain network structures and capture dependencies between connected nodes representing different locations or echelons [33].

Inventory optimization in multi-echelon supply chains traditionally relies on mathematical programming formulations and analytical models that determine optimal order quantities, reorder points, and safety stock levels across network tiers [34]. Classic approaches include echelon stock policies, installation stock policies, and base-stock policies that prescribe inventory positioning strategies based on demand characteristics and cost structures [35]. Stochastic inventory models incorporate demand uncertainty through probability distributions and optimize inventory decisions under service level constraints or cost minimization objectives [36]. However, these traditional models typically assume stationary demand distributions and may not fully leverage the rich probabilistic information provided by modern forecasting approaches [37].

Recent research has explored the integration of DL forecasts with inventory optimization frameworks through various methodologies [38]. Prescriptive analytics approaches that directly optimize decision variables using forecasting models as components within optimization algorithms have demonstrated improved performance over sequential forecast-then-optimize procedures [39]. End-to-end learning frameworks that jointly train forecasting and decision-making modules enable the forecasting model to learn patterns most relevant for downstream optimization objectives rather than purely minimizing forecast errors [40]. Reinforcement learning (RL) combined with transformer-based forecasting has been applied to develop adaptive inventory policies that learn optimal actions through interaction with supply chain environments [41]. These hybrid approaches demonstrate the potential for tighter integration between prediction and prescription in supply chain decision-making.



Key Characteristics Comparison

Characteristic	Gen 1	Gen 2	Gen 3	Gen 4
Long-Range Dependencies:	● Low	● Medium	● Medium-High	● High
Parallel Processing:	✗ No	▲ Partial	▲ Limited	✓ Yes
Computational Complexity:	O(n)	O(n log n)	O(n²)	O(n log n)
Feature Engineering:	● Manual	● Required	✓ Automatic	✓ Automatic

Figure 1. A comparative flowchart showing the evolution of demand forecasting methods in supply chains, from traditional statistical methods (ARIMA, Exponential Smoothing) to ML approaches (Random Forests, SVMs) to DL architectures (LSTM, GRU, CNN) and finally to Transformer-based models. The figure should illustrate key characteristics of each generation including computational complexity, ability to capture long-range dependencies, and parallel processing capability.

3. Transformer Architectures for Supply Chain Demand Forecasting

Transformer architectures designed for supply chain demand forecasting must address several domain-specific challenges including irregular sampling intervals, missing data, incorporation of categorical variables, and handling of multiple related time series simultaneously [42]. The core self-attention mechanism computes attention weights between all pairs of time steps in the input sequence, enabling the model to identify relevant historical patterns regardless of their temporal distance from the prediction point [43]. This formulation allows the model to adaptively focus on different historical periods depending on the current context and prediction requirements. Positional encoding represents a critical component of transformer architectures, providing the model with information about the temporal ordering of input elements [44]. For time series applications, traditional sinusoidal positional encodings may be augmented or replaced with learned embeddings that capture calendar effects, seasonal patterns, and other temporal regularities specific to supply chain demand [45].

The multi-head attention mechanism extends the basic attention operation by computing multiple parallel attention functions with different learned transformations, enabling the model to attend to information from different representation subspaces [46]. In supply chain forecasting contexts, different attention heads may specialize in capturing distinct patterns such as trend components, seasonal variations, promotional effects, or external factor influences [47]. Analysis of attention weights across heads provides interpretability regarding which temporal patterns and input features drive predictions for specific products or time periods [48]. Encoder-decoder transformer architectures separate the processing of historical observations from the generation of future predictions, with cross-attention mechanisms enabling the decoder to selectively attend to relevant encoder outputs [49]. This structure proves particularly suitable for multi-horizon forecasting problems common in supply chain planning, where predictions are required for multiple future time steps simultaneously [50].

Sparse attention mechanisms address computational and memory limitations of standard self-attention when processing very long sequences, which is common in supply chain scenarios with years of historical data [51]. The ProbSparse attention employed in Informer selectively computes attention only for the most informative query-key pairs based on a sparsity measurement, reducing computational complexity from quadratic to logarithmic with respect to sequence length [52]. Variable selection mechanisms integrated within transformer architectures enable the model to identify and emphasize the most relevant input features for demand forecasting [53]. Static covariates such as product categories and location characteristics typically remain constant over time and can be processed through separate encoding pathways before integration with temporal features, while time-varying covariates including prices, promotional indicators, and economic indicators exhibit temporal dynamics that must be jointly modeled with historical demand patterns [54].

Uncertainty quantification represents a crucial requirement for supply chain applications, as inventory optimization and risk management decisions depend critically on understanding prediction confidence and potential forecast errors [55]. Probabilistic forecasting extensions of transformer architectures generate complete predictive distributions rather than single-point estimates, enabling calculation of prediction intervals and risk metrics [56]. Quantile regression approaches train the model to predict multiple quantiles of the forecast distribution simultaneously, providing non-parametric uncertainty estimates that do not assume specific distributional forms [57].

Table 1. Comparative performance metrics of transformer-based forecasting models versus traditional methods for supply chain demand prediction. The table should include columns for Model Type, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Computational Time. Rows should compare ARIMA, LSTM, Temporal Fusion Transformer, Informer, and Autoformer across retail and manufacturing supply chain datasets.

Retail Supply Chain				
Model	**MAE**	**RMSE**	**MAPE (%)**	**Time (sec)**
ARIMA	45.3	62.8	18.5	2.3
LSTM	32.7	48.4	13.2	15.7
TFT	24.1	36.9	9.8	28.4
Informer	25.8	38.2	10.3	18.6
Autoformer	23.6	35.4	9.5	22.1
Manufacturing Supply Chain				
Model	MAE	RMSE	MAPE (%)	Time (sec)
ARIMA	128.5	187.3	22.4	1.8
LSTM	96.2	142.6	16.8	12.4
TFT	68.4	102.7	11.9	24.8
Informer	72.1	108.4	12.6	16.2
Autoformer	66.8	100.3	11.4	19.7
Distribution Network				
Model	MAE	RMSE	MAPE (%)	Time (sec)
ARIMA	87.6	124.5	20.7	2.1
LSTM	64.3	95.8	15.3	14.2
TFT	46.7	71.2	11.2	26.9
Informer	49.2	74.6	11.8	17.8
Autoformer	45.3	69.7	10.8	20.5

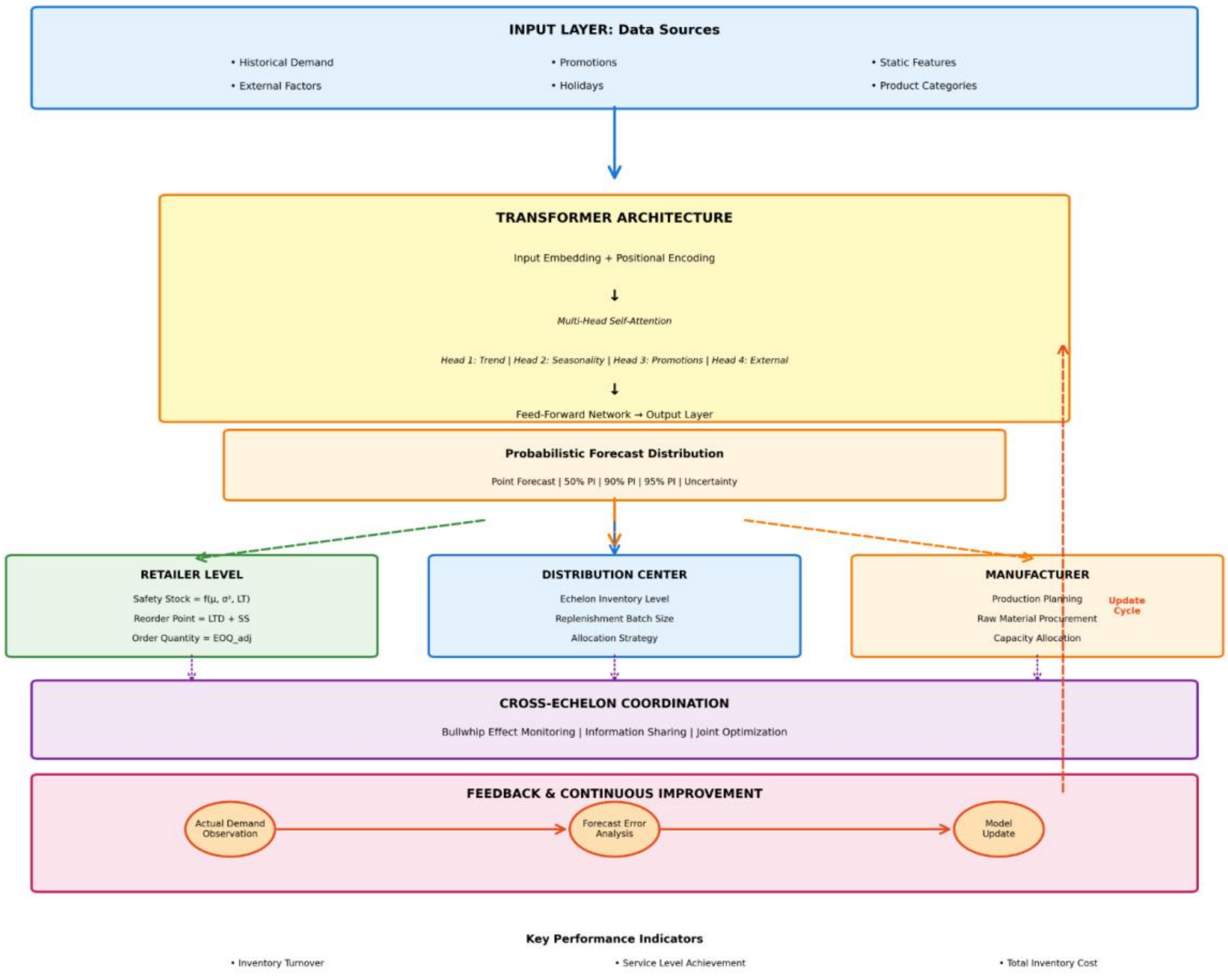
4. Integration of Transformer Forecasts with Inventory Optimization

The integration of transformer-based demand forecasts with inventory optimization requires careful consideration of how probabilistic predictions inform inventory decision variables across multiple echelons [58]. Traditional inventory models optimize order quantities and reorder points based on demand distributions characterized by mean and variance parameters, which may inadequately represent the rich predictive information available from transformer models [59]. Advanced integration approaches extract relevant statistics from transformer-generated forecast distributions, including time-varying means, prediction intervals, and tail risk measures, to parameterize adaptive inventory policies [60]. This integration enables inventory decisions to respond dynamically to changing demand patterns and forecast uncertainty rather than relying on static parameters estimated from historical data.

Safety stock calculations in multi-echelon supply chains critically depend on demand variability and lead time uncertainty, both of which can be more accurately characterized using transformer forecasts. The traditional safety stock formula can be extended to incorporate time-varying uncertainty estimates provided by transformer models, adjusting buffer inventory levels as forecast confidence changes [61]. Prediction intervals generated by probabilistic transformer models directly inform safety stock requirements by specifying the inventory level needed to achieve desired service levels under forecast uncertainty. Dynamic safety stock policies that adapt to seasonal demand patterns and promotional events can be derived from transformer forecasts that capture these effects [62].

Reorder point determination in periodic review inventory systems leverages transformer forecasts to estimate demand during lead time plus review period intervals. Multi-step ahead transformer predictions provide natural inputs for reorder point calculations in supply chains with substantial lead times, avoiding the need for simplified assumptions about demand stationarity. The bullwhip effect mitigation achieved through improved forecasting accuracy at downstream echelons propagates benefits throughout the supply chain network, reducing inventory variability and improving service levels at upstream stages [63]. Coordinated inventory policies that explicitly model dependencies between echelons can incorporate transformer forecasts at multiple network locations simultaneously, optimizing system-wide inventory positions while accounting for lead time interdependencies [64].

Order quantity optimization traditionally employs economic order quantity formulas or dynamic lot-sizing algorithms that balance ordering costs against holding costs. Integration with transformer forecasts enables more responsive ordering policies that adjust batch sizes based on anticipated demand patterns and forecast confidence levels. When transformer models predict high demand periods with high confidence, order quantities can be increased to ensure adequate inventory availability, while periods of uncertain or declining demand may warrant reduced order sizes to minimize holding costs and obsolescence risks. Multi-product inventory optimization problems that consider substitution effects, capacity constraints, and shared resources benefit from transformer models' ability to forecast demand for multiple related products simultaneously while capturing cross-product dependencies through attention mechanisms.


INPUT LAYER: Data Sources
• Historical Demand
• External Factors
• Promotions
• Holidays
• Static Features
• Product Categories
TRANSFORMER ARCHITECTURE
Input Embedding + Positional Encoding
Multi-Head Self-Attention
Head 1: Trend | Head 2: Seasonality | Head 3: Promotions | Head 4: External
Feed-Forward Network → Output Layer
Probabilistic Forecast Distribution
Point Forecast | 50% PI | 90% PI | 95% PI | Uncertainty
RETAILER LEVEL
Safety Stock = $f(\mu, \sigma^2, LT)$
Reorder Point = LTD + SS
Order Quantity = EOQ_adj
DISTRIBUTION CENTER
Echelon Inventory Level
Replenishment Batch Size
Allocation Strategy
MANUFACTURER
Production Planning
Raw Material Procurement
Capacity Allocation
Update Cycle
CROSS-ECHELON COORDINATION
Bullwhip Effect Monitoring | Information Sharing | Joint Optimization
FEEDBACK & CONTINUOUS IMPROVEMENT
Actual Demand Observation
Forecast Error Analysis
Model Update
Key Performance Indicators
• Inventory Turnover
• Service Level Achievement
• Total Inventory Cost

Architecture based on: Oroojlooyjadid et al. (2022), Lim et al. (2021)

Figure 2. An integrated framework diagram illustrating the connection between transformer-based demand forecasting and multi-echelon inventory optimization. The diagram should show data flow from historical demand and external factors through transformer architecture components (positional encoding, multi-head attention, feed-forward layers) to generate probabilistic forecasts, which then inform inventory decision variables (safety stocks, reorder points, order quantities) at different echelon levels (retailer, distribution center, manufacturer). The framework should highlight feedback loops and coordination mechanisms across echelons.

5. Performance Analysis and Implementation Considerations

Empirical evaluations of transformer-based forecasting methods in supply chain contexts have demonstrated substantial performance improvements over traditional approaches across multiple metrics and application domains. Studies comparing transformer architectures against ARIMA, LSTM, and other baseline methods consistently report reductions in mean absolute percentage error ranging from fifteen to thirty-five percent depending on dataset characteristics and forecast horizons [65]. The performance advantages of transformers become particularly pronounced for long-horizon forecasts spanning multiple weeks or months, where the ability to capture long-range dependencies and seasonal patterns proves crucial [66]. Retail demand forecasting applications have shown that transformer models effectively handle promotional effects, holiday seasonality, and trend changes that challenge conventional methods [67].

Computational considerations represent an important practical factor in deploying transformer models for supply chain forecasting at scale. While transformers enable parallel processing of sequences, standard self-attention mechanisms exhibit quadratic computational complexity with respect to sequence length, potentially limiting applicability for very long historical windows [68]. Sparse attention variants and efficient transformer architectures specifically designed for long sequences address these scalability challenges, enabling processing of multi-year historical datasets with manageable computational resources [69]. Hardware acceleration through graphics processing units and specialized tensor processing units further enhances training and inference efficiency, making real-time forecasting feasible even for large product portfolios.

Data quality and preprocessing requirements significantly influence transformer model performance in supply chain applications. Missing values, outliers, and irregular sampling intervals necessitate careful data preparation strategies [70]. Imputation methods for handling missing observations should preserve temporal patterns and avoid introducing artificial smoothness that may degrade forecast quality. Outlier detection and treatment procedures must distinguish between genuine demand spikes driven by promotions or external events versus data errors requiring correction. Feature engineering and selection processes identify relevant external variables and construct informative input representations, though transformer architectures' inherent feature learning capabilities reduce the extent of manual feature engineering compared to traditional ML approaches.

Model training procedures for supply chain forecasting transformers involve several important design choices regarding loss functions, optimization algorithms, and regularization techniques. Mean squared error and mean absolute error represent common loss functions for point forecasting, while quantile loss and maximum likelihood objectives enable probabilistic forecasting [71]. Appropriate training data splitting strategies that respect temporal ordering and avoid data leakage ensure valid performance evaluation. Cross-validation approaches adapted for time series data, such as rolling window validation and expanding window validation, provide robust estimates of generalization performance across different forecast horizons and demand patterns. Hyperparameter tuning through systematic grid search or Bayesian optimization identifies effective configurations for model capacity, learning rates, and regularization strengths.

Model interpretability and explainability constitute critical requirements for supply chain practitioners seeking to understand and trust transformer forecasts. Attention weight visualization reveals which historical time periods most strongly influence specific predictions, providing insights into demand drivers and seasonal patterns [72]. Feature importance analysis quantifies the contribution of different input variables to forecast accuracy, guiding data collection priorities and identifying key demand determinants. Counterfactual explanations that show how predictions would change under alternative scenarios support what-if analysis and scenario planning activities common in supply chain decision-making contexts.

6. Emerging Trends and Future Directions

Recent developments in transformer-based supply chain forecasting explore several promising directions that address current limitations and expand application domains. Federated learning (FL) approaches enable collaborative forecasting across multiple organizations while preserving data privacy and competitive sensitivities [73]. In this paradigm, individual supply chain partners train local transformer models on proprietary data and share only model updates rather than raw data, enabling the development of more accurate forecasts that leverage broader information while respecting confidentiality constraints. FL proves particularly valuable for supply chain networks involving multiple independent entities that could benefit from coordinated forecasting but face barriers to direct data sharing.

Hybrid architectures combining transformers with other neural network types or algorithmic components represent another active research frontier. Transformer-GNN hybrids model both temporal dynamics and spatial relationships within supply chain networks, capturing how demand patterns propagate across connected locations and echelons [74]. Integration of transformers with physics-based models or causal reasoning frameworks enables incorporation of domain knowledge and structural constraints into forecasting systems, potentially improving generalization to novel situations not well-represented in historical data. Transformer models combined with RL create adaptive inventory control systems that learn optimal policies through simulation or real-world interaction while leveraging transformer forecasts as state representations.

Transfer learning and pre-training strategies adapted for supply chain forecasting enable knowledge transfer across products, locations, or organizations [75]. Large-scale pre-training on diverse supply chain datasets creates foundation models that capture general temporal patterns and demand dynamics, which can then be fine-tuned for specific forecasting tasks with limited data. This approach proves especially valuable for new product introductions, market expansions, or supply chain network reconfigurations where historical data may be scarce. Domain adaptation techniques address distribution shifts between pre-training and target domains, ensuring effective transfer despite differences in demand characteristics.

Attention mechanism innovations continue to enhance transformer capabilities for supply chain applications. Causal attention mechanisms that explicitly model cause-effect relationships between input features and demand outcomes improve interpretability and support intervention analysis [76]. Multi-resolution attention architectures that simultaneously capture patterns at different temporal scales enable more effective modeling of supply chains exhibiting nested seasonality and hierarchical time structures. Adaptive attention mechanisms that dynamically adjust computational allocation based on input complexity optimize the trade-off between forecast accuracy and computational efficiency.

Integration with broader supply chain decision support systems represents a crucial direction for practical impact. Transformer forecasts should seamlessly feed into enterprise resource planning systems, warehouse management systems, and transportation planning tools [77]. Real-time forecast updating as new demand observations arrive enables responsive supply chain management that quickly adapts to changing conditions. Scenario-based forecasting capabilities that generate predictions under alternative assumptions about promotions, competitor actions, or market conditions support strategic planning and risk management activities.

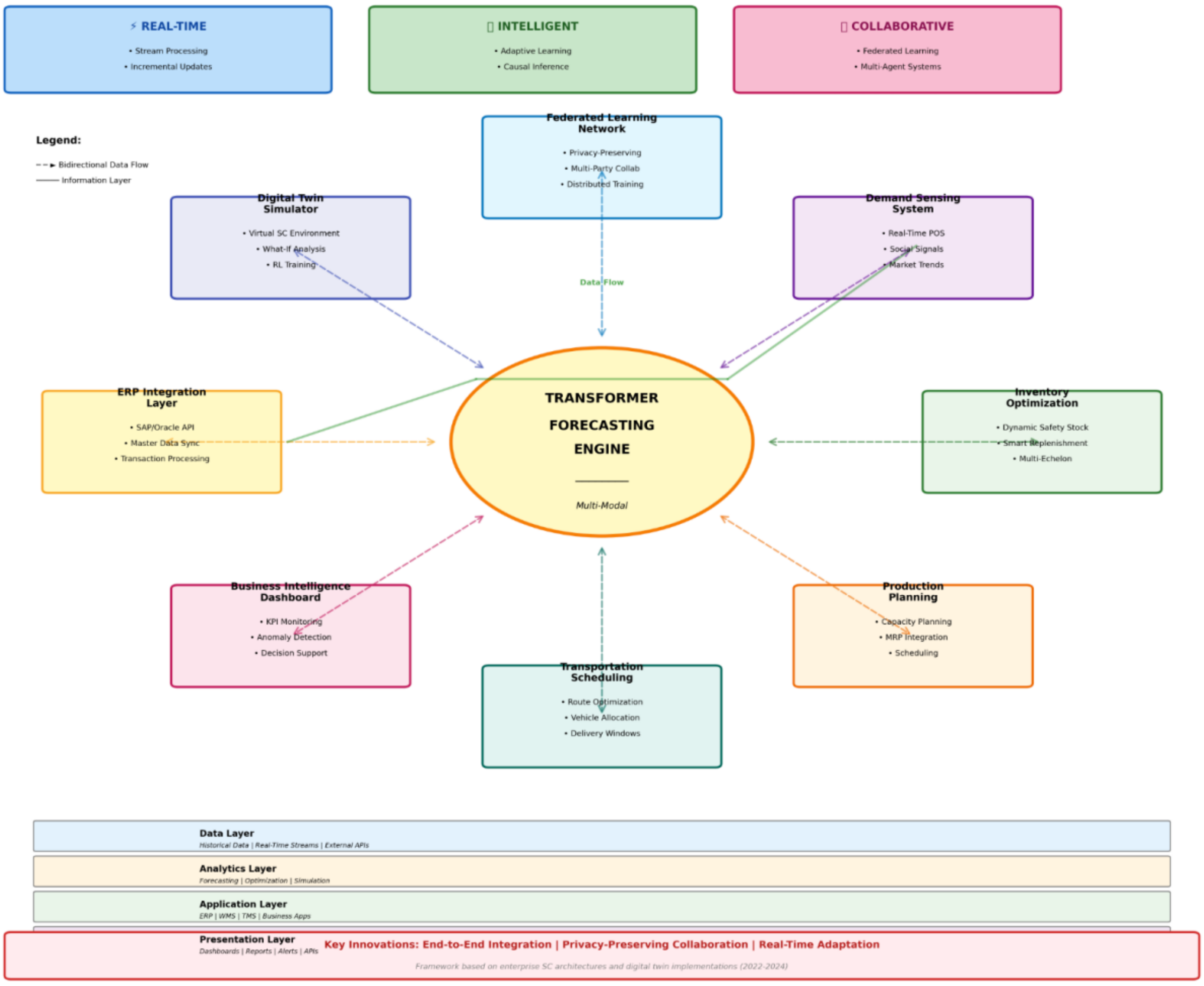


Figure 3. A future-oriented schematic showing the integration of transformer-based forecasting within an end-to-end supply chain decision support ecosystem. The diagram should illustrate connections between transformer forecasting models and various supply chain management components including demand sensing systems, inventory optimization engines, production planning modules, transportation scheduling systems, and business intelligence dashboards. The schematic should highlight emerging capabilities such as federated learning across supply chain partners, real-time forecast updating, and integrated prescriptive analytics.

7. Conclusion

Transformer-based architectures represent a significant advancement in demand forecasting capabilities for multi-echelon supply chain networks, offering substantial improvements over traditional statistical methods and earlier DL approaches. The self-attention mechanisms underlying transformer models enable effective capture of long-range temporal dependencies, complex seasonal patterns, and multivariate relationships that characterize supply chain demand. Specialized transformer variants designed for time series applications address computational scalability challenges and incorporate domain-specific modifications that enhance forecasting accuracy for supply chain contexts. Empirical evaluations across diverse supply chain settings consistently demonstrate performance advantages measured through reduced forecast errors and improved inventory optimization outcomes.

The integration of transformer forecasts with inventory optimization frameworks enables more responsive and adaptive supply chain management. Probabilistic forecasts generated by transformer models inform dynamic safety stock calculations, responsive reorder point determination, and flexible order quantity decisions that adapt to changing demand patterns and forecast uncertainty. Multi-echelon coordination mechanisms that leverage transformer predictions across supply chain tiers mitigate bullwhip effect amplification and improve system-wide inventory positioning. End-to-end learning approaches that jointly optimize forecasting and decision-making components demonstrate the potential for tighter integration between prediction and prescription in supply chain planning.

Several important challenges and research opportunities remain in advancing transformer-based supply chain forecasting. Scalability to very large product portfolios and extended historical windows requires continued development of efficient attention mechanisms and computational optimization strategies. Robustness to demand distribution shifts, unprecedented events, and structural changes in supply chain networks necessitates techniques for handling non-stationarity and detecting concept drift. Integration of causal reasoning and domain knowledge into transformer architectures may improve generalization beyond patterns observed in historical data. Collaborative forecasting frameworks that enable knowledge sharing while respecting competitive sensitivities represent important directions for supply chain networks involving multiple independent organizations.

The interpretability and explainability of transformer forecasts remain critical for practitioner adoption and trust. Attention visualization techniques provide valuable insights into demand drivers and temporal patterns, but further work is needed to translate model explanations into actionable supply chain insights. Uncertainty quantification methods that accurately characterize forecast confidence and tail risks require continued refinement, particularly for rare but impactful demand scenarios. Integration with broader supply chain decision support systems and enterprise software platforms will facilitate practical deployment and realize the full potential of transformer-based forecasting in operational settings.

Future developments in transformer architectures specifically designed for supply chain applications promise continued advances in forecasting accuracy, computational efficiency, and practical utility. Hybrid models combining transformers with complementary approaches, transfer learning strategies enabling knowledge reuse across domains, and adaptive mechanisms that respond to changing conditions represent particularly promising directions. As organizations increasingly adopt data-driven supply chain management practices, transformer-based forecasting will play a central role in enabling responsive, efficient, and resilient supply chain networks capable of thriving in complex and uncertain environments. The ongoing evolution of these technologies, coupled with growing availability of supply chain data and computational resources, positions transformer-based approaches as foundational tools for next-generation supply chain analytics and optimization.

References

Dolgui, A., Ivanov, D., & Sokolov, B. (2020). Reconfigurable supply chain: The X-network. *International Journal of Production Research, 58*(13), 4138–4163.

Abolghasemi, M., Rostami-Tabar, B., & Syntetos, A. (2023). The value of point-of-sales information in upstream supply chain forecasting: An empirical investigation. *International Journal of Production Research, 61*(7), 2162–2177.

Kilimci, Z. H., Akyuz, A. O., Uysal, M., et al. (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. *Complexity, 2019*, 9067367.

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A, 379*(2194), 20200209.

Wen, Q., Zhou, T., Zhang, C., et al. (2022). Transformers in time series: A survey. *arXiv Preprint arXiv:2202.07125.*

Roy, A., Saffar, M., Vaswani, A., & Grangier, D. (2021). Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics, 9*, 53–69.

Wang, S., Lin, Y., Jia, Y., Sun, J., & Yang, Z. (2024). Unveiling the multi-dimensional spatio-temporal fusion transformer (MDSTFT): A revolutionary deep learning framework for enhanced multivariate time series forecasting. *IEEE Access.*

Sánchez-Durán, R., Luque, J., & Barbancho, J. (2019). Long-term demand forecasting in a scenario of energy transition. *Energies, 12*(16), 3095.

Bogale, M., & Desta, E. (2025). Mitigating the bullwhip effect through sustainable supply chain practices: A systematic literature review. *Ethiopian Journal of Development Research, 47*(1), 34–56.

Punia, S., Singh, S. P., & Madaan, J. K. (2020). From predictive to prescriptive analytics: A data-driven multi-item newsvendor model. *Decision Support Systems, 136*, 113340.

Garuba, A. A., Akanni, G. O., & Adebayo, A. M. (2025). Application of time series analysis in management sciences research: Trend, forecasting and decision making. *Bayero Journal of Social Science and Administration, 10*(1), 44–73.

Bougioukos, V., Nikolopoulos, K., & Tsinopoulos, C. (2025). Introduction to forecasting, planning and strategy in a turbulent era. In *Forecasting, planning and strategy in a turbulent era* (pp. 1–6). Edward Elgar Publishing.

Feizabadi, J. (2022). Machine learning demand forecasting and supply chain performance. *International Journal of Logistics Research and Applications, 25*(2), 119–142.

Sakib, M., Mustajab, S., & Alam, M. (2025). Ensemble deep learning techniques for time series analysis: A comprehensive review, applications, open issues, challenges, and future directions. *Cluster Computing, 28*(1), 73.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., ... & Hochreiter, S. (2024). XLSTM: Extended long short-term memory. *Advances in Neural Information Processing Systems, 37*, 107547–107603.

Wickramasuriya, S., Bandara, K., Hewamalage, H., & Perera, M. (2024). Forecasting hierarchical time series using non-linear mappings. *SSRN.* https://ssrn.com/abstract=4793559

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., ... & Xie, L. (2020). DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv Preprint arXiv:2008.00264.*

Cheng, Y., Liu, Z., & Morimoto, Y. (2020). Attention-based SeriesNet: An attention-based hybrid neural network model for conditional time series forecasting. *Information, 11*(6), 305.

Wang, Y., Ding, G., Zeng, Z., & Yang, S. (2025). Causal-aware multimodal transformer for supply chain demand forecasting: Integrating text, time series, and satellite imagery. *IEEE Access.*

Zeng, A., Chen, M., Zhang, L., et al. (2023). Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(9), 11121–11128.

Lim, B., Arık, S. Ö., Loeff, N., et al. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting, 37*(4), 1748–1764.

Kamarthi, H., Kong, L., Rodríguez, A., et al. (2022). CAMul: Calibrated and accurate multi-view time-series forecasting. In *Proceedings of the ACM Web Conference* (pp. 3174–3185).

Zhu, Q., Han, J., Chai, K., & Zhao, C. (2023). Time series analysis based on informer algorithms: A survey. *Symmetry, 15*(4), 951.

Su, G., & Guan, Y. (2025). MSDformer: An autocorrelation transformer with multiscale decomposition for long-term multivariate time series forecasting. *Applied Intelligence, 55*(3), 179.

Liu, S., Yu, H., Liao, C., et al. (2022). Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. *International Conference on Learning Representations.*

Bian, J., Al Arafat, A., Xiong, H., Li, J., Li, L., Chen, H., ... & Guo, Z. (2022). Machine learning in real-time Internet of Things (IoT) systems: A survey. *IEEE Internet of Things Journal, 9*(11), 8364–8386.

Shen, Y., Yang, X., Zhao, J., & Li, Z. (2023, December). An analysis of combined data augmentation in time series prediction tasks using discrete wavelet multilevel decomposition. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 366–371). IEEE.

Salinas, D., Flunkert, V., Gasthaus, J., et al. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting, 36*(3), 1181–1191.

Wu, N., Green, B., Ben, X., et al. (2020). Deep transformer models for time series forecasting: The influenza prevalence case. *arXiv Preprint arXiv:2001.08317.*

Zhang, M., Saab, K. K., Poli, M., Dao, T., Goel, K., & Ré, C. (2023). Effectively modeling time series with simple discrete state spaces. *arXiv Preprint arXiv:2303.09489.*

Tan, Z., He, S., Zhan, H., Huang, Y., & Huang, S. (2025, March). Graphormer-based Bayesian network conditional normalizing flow for multivariate time series anomaly detection in communication networks. In *2025 IEEE Wireless Communications and Networking Conference (WCNC)* (pp. 1–6). IEEE.

Agarwal, K., Dheekollu, L., Dhama, G., Arora, A., Asthana, S., & Bhowmik, T. (2021). Deep learning-based time series forecasting. In *Deep learning applications (Vol. 3)* (pp. 151–169). Springer Singapore.

Cao, D., Wang, Y., Duan, J., et al. (2020). Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in Neural Information Processing Systems, 33*, 17766–17778.

Paldrak, M., Erol, E., İnan, A., Fırat, D., Miran, A. E., Çetinkaya, E., ... & Aydın, P. (2022, October). Demand forecasting and inventory control system for industrial valves. In *International Symposium for Production Research* (pp. 780–796). Springer International Publishing.

Shen, X., Yu, Y., & Song, J. S. (2022). Optimal policies for a multi-echelon inventory problem with service time target and expediting. *Manufacturing & Service Operations Management, 24*(4), 2310–2327.

Sezer, M., & Çebi, F. (2024). Comparative analysis of probabilistic models for intermittent demand forecasting. In *Engineering and technology management in challenging times* (pp. 159–171). Springer Nature Switzerland.

Babai, M. Z., Boylan, J. E., & Rostami-Tabar, B. (2022). Demand forecasting in supply chains: A review of aggregation and hierarchical approaches. *International Journal of Production Research, 60*(1), 324–348.

Theodoridis, G., & Tsadiras, A. (2024). Retail demand forecasting: A multivariate approach and comparison of boosting and deep learning methods. *International Journal on Artificial Intelligence Tools, 33*(4), 2450001.

Notz, P. M., & Pibernik, R. (2022). Prescriptive analytics for flexible capacity management. *Management Science, 68*(3), 1756–1775.

Wen, J., Abeel, T., & de Weerdt, M. (2025). Performance and interaction assessment of neural network architectures and bivariate smart predict-then-optimize. *Machine Learning, 114*(11), 252.

Oroojlooyjadid, A., Snyder, L. V., & Takáč, M. (2020). Applying deep learning to the newsvendor problem. *IISE Transactions, 52*(4), 444–463.

Alexandrov, A., Benidis, K., Bohlke-Schneider, M., et al. (2020). GluonTS: Probabilistic and neural time series modeling in Python. *Journal of Machine Learning Research, 21*(116), 1–6.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022, June). FedFormer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning* (pp. 27268–27286). PMLR.

Kazemi, S. M., Goel, R., Eghbali, S., et al. (2019). Time2Vec: Learning a vector representation of time. *arXiv Preprint arXiv:1907.05321.*

Zhou, H., Zhang, S., Peng, J., et al. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(12), 11106–11115.

Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020, November). Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning* (pp. 5156–5165). PMLR.

Liu, Y., Wu, H., Wang, J., et al. (2022). Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems, 35*, 9881–9893.

Zhang, Y., & Yan, J. (2023). Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. *International Conference on Learning Representations.*

Ding, C., Wang, G., Zhang, X., Liu, Q., & Liu, X. (2021). A hybrid CNN-LSTM model for predicting PM2.5 in Beijing based on spatiotemporal correlation. *Environmental and Ecological Statistics, 28*(3), 503–522.

Hyndman, R. J. (2023). Forecasting, causality and feedback. *International Journal of Forecasting, 39*(2), 558–560.

Child, R., Gray, S., Radford, A., et al. (2019). Generating long sequences with sparse transformers. *arXiv Preprint arXiv:1904.10509.*

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv Preprint arXiv:2004.05150.*

Herzen, J., Lässig, F., Piazzetta, S. G., et al. (2022). Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research, 23*(124), 1–6.

Fan, C., Zhang, Y., Pan, Y., Li, X., Zhang, C., Yuan, R., ... & Huang, H. (2019, July). Multi-horizon time series forecasting with temporal attention learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2527–2535).

Mobiny, A., Yuan, P., Moulik, S. K., Garg, N., Wu, C. C., & Van Nguyen, H. (2021). DropConnect is effective in modeling uncertainty of Bayesian deep networks. *Scientific Reports, 11*(1), 5458.

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software, 90*, 1–37.

Chernozhukov, V., Fernández-Val, I., & Melly, B. (2022). Fast algorithms for the quantile regression process. *Empirical Economics, 62*(1), 7–33.

Punia, S., Nikolopoulos, K., Singh, S. P., et al. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *International Journal of Production Research, 58*(16), 4964–4979.

Ivanov, D., Tsipoulanidis, A., & Schönberger, J. (2021). Inventory management. In *Global supply chain and operations management: A decision-oriented introduction to the creation of value* (pp. 385–433). Springer International Publishing.

Zhang, D., Pee, L. G., & Cui, L. (2021). Artificial intelligence in e-commerce fulfillment: A case study of resource orchestration at Alibaba's smart warehouse. *International Journal of Information Management, 57*, 102304.

Althaqafi, T. (2024). A study on inventory control system for a supply chain using Markov decision processes. *Edelweiss Applied Science and Technology, 8*(6), 7846–7864.

Chen, L., Dong, T., Peng, J., & Ralescu, D. (2023). Uncertainty analysis and optimization modeling with application to supply chain management: A systematic review. *Mathematics, 11*(11), 2530.

Lopez-Campos, M., Cannella, S., Miranda, P. A., & Stegmaier, R. (2019). Modeling the operation of synchronized supply chains under a collaborative structure. *Academia Revista Latinoamericana de Administración, 32*(2), 203–224.

De Amadi, C. (2024). Influence of lean supply chain management strategies and organizational sustainability of manufacturing firms in Rivers State.

Weng, T., Liu, W., & Xiao, J. (2020). Supply chain sales forecasting based on LightGBM and LSTM combination model. *Industrial Management & Data Systems, 120*(2), 265–279.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting, 38*(4), 1346–1364.

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting, 37*(2), 587–603.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys, 54*(10s), 1–41.

Zaheer, M., Guruganesh, G., Dubey, K. A., et al. (2020). Big Bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems, 33*, 17283–17297.

Barak, S., Nasiri, M., & Rostamzadeh, M. (2019). Time series model selection with a meta-learning approach: Evidence from a pool of forecasting algorithms. *arXiv Preprint arXiv:1908.08489.*

Wang, R., Yan, F., Shi, R., Yu, L., & Deng, Y. (2022). Uncertainty-controlled remaining useful life prediction of bearings with a new data-augmentation strategy. *Applied Sciences, 12*(21), 11086.

Sun, K., Wang, L., Xu, B., Zhao, W., Teng, S. W., & Xia, F. (2020). Network representation learning: From traditional feature learning to deep learning. *IEEE Access, 8*, 205600–205617.

Yang, Q., Liu, Y., Chen, T., et al. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology, 10*(2), 1–19.

Teng, J., Jiang, Y., & Shi, R. (2024). GraphDeformer: A spatio-temporal model integrating graph neural network and transformer for wind power forecasting. *SSRN.* https://ssrn.com/abstract=4750487

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43–76.

Schölkopf, B. (2022). Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl* (pp. 765–804).

Ivanov, D., Dolgui, A., & Sokolov, B. (2019). The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics. *International Journal of Production Research, 57*(3), 829–846.