

Article

Beyond Traditional Forecasting Methods: Evaluating LSTM Performance on Diverse Time Series

Zoltán Baráth, Péter Veres  and Ágota Banyai * 

Institute of Logistics, University of Miskolc, 3515 Miskolc, Hungary; zoltan.barath@student.uni-miskolc.hu (Z.B.); peter.veres@uni-miskolc.hu (P.V.)

* Correspondence: agota.banyaine@uni-miskolc.hu

Abstract

Time series forecasting performance is strongly influenced by the structural properties of the underlying data, yet learning-based models are often applied without sufficient validation of this dependency. This study evaluates a uniformly configured Long Short-Term Memory (LSTM) model on five real-world weekly time series with different levels of periodicity, noise, and volatility. Forecasting is performed in a single-step setting using a fixed sliding window of 12 weeks under a consistent training, validation, and testing framework. Model performance is assessed using mean squared error (MSE) and the coefficient of determination R^2 . The results show that for well-structured series, both the LSTM model and Holt's exponential smoothing achieve very low MSE values with R^2 scores close to one, indicating excellent predictive accuracy. For other items, performance varies across methods, with either the LSTM or Holt model providing the best results depending on the data structure. These findings confirm that high forecasting accuracy can be achieved with both advanced and classical methods, and that data characteristics play a more decisive role than model complexity.

Keywords: LSTM; time series forecasting; forecast accuracy; statistical forecasting; ARIMA

MSC: 68T07; 62M10; 62P30

1. Introduction and Background

Forecasting has been used for centuries to support decisions in everyday life, such as planning agriculture, managing trade, or predicting weather conditions. Early forecasting relied on observation and experience, and later developed into statistical models that describe trends, cycles, and uncertainty. Major milestones included the formalization of probability theory and the introduction of time series models such as autoregressive [1] and moving average processes [2], which enabled uncertainty to be quantified and future values to be estimated in a systematic manner. These advances established forecasting as a scientific discipline and laid the foundation for modern predictive methods used across many domains.

Artificial intelligence emerged as a much later development, driven by the ambition to construct systems capable of learning from data rather than relying solely on predefined rules. Early approaches focused on symbolic reasoning systems, while subsequent advances introduced machine learning algorithms such as decision trees and neural networks that adapt their behavior through experience. With the availability of large datasets and increased computational capacity, these models have become effective tools for identifying



Academic Editor: Pedro Carpena

Received: 19 January 2026

Revised: 19 February 2026

Accepted: 27 February 2026

Published: 1 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

complex patterns and nonlinear relationships. As a result, artificial intelligence has gradually become a natural extension of forecasting, offering new possibilities for prediction while also introducing new sources of uncertainty and methodological challenges.

1.1. Forecasting Importance in Decision-Making Systems

Forecasts are not merely descriptive summaries of historical patterns; in most operational settings, they function as inputs to decisions. In supply chains, production planning, procurement, and inventory control, forecasts directly influence order quantities, capacity plans, replenishment schedules, and service-level commitments. Consequently, forecast errors propagate through planning systems and can amplify upstream decisions, increasing costs or deteriorating customer service. A classical illustration of such amplification is the bullwhip effect: when demand information is transmitted through orders, variability and distortion tend to grow along the supply chain, affecting inventory and production decisions of upstream members [3]. This systemic role implies that the value of forecasting cannot be assessed only by statistical fit, but by its impact on downstream decisions and operational performance.

From an inventory perspective, forecast uncertainty is a primary driver of safety stock and service levels. When demand is forecasted rather than assumed known, decision rules need to explicitly account for forecast error; otherwise, inventory buffers can be systematically under- or over-dimensioned. Research on safety-stock calculation under forecasted demand highlights that ignoring the structure of forecast errors (e.g., correlation over time, differences between forecast interval and lead time) can lead to biased service-level performance and suboptimal inventory investments [4]. This connection between forecasting and inventory performance is especially salient in weekly operational planning, where short cycles leave limited time for corrective actions and where forecast errors quickly translate into material shortages, expedited shipments, or excess stock.

In practice, the economic consequences of forecast errors are often asymmetric. Over-forecasting may inflate holding costs and obsolescence risk, whereas under-forecasting can trigger stockouts, lost sales, or costly expediting. Because these costs differ by item criticality, lead time, and substitution possibilities, the same numerical error can imply very different operational outcomes. This asymmetry motivates forecasting processes that are tailored to decision contexts (e.g., procurement vs. finished goods planning) and underscores that improving accuracy is not an abstract academic objective, but a means to reduce operational risk. Practitioner-oriented frameworks in demand planning emphasize that forecast accuracy targets and evaluation must be aligned to the horizon and use case, rather than treated as universal benchmarks [5].

At the same time, the forecasting landscape has been transformed by the widespread adoption of machine learning and deep learning methods, raising expectations of “automatic” accuracy gains. Large-scale empirical evidence, however, suggests that progress is nuanced: competitions and benchmarks demonstrate that performance depends strongly on data characteristics and that robust baselines and model combinations often remain hard to beat across heterogeneous time series collections [6]. This is consistent with the broader forecasting literature, which stresses that practical gains come as much from process quality (appropriate evaluation design, leakage-free validation, and careful metric selection) as from model class [7]. In other words, even sophisticated models can appear successful under weak evaluation protocols, while well-designed evaluations frequently reveal that accuracy improvements are conditional rather than universal.

Organizational practice further reinforces why forecasting “importance” should be discussed beyond model mechanics. In many firms, forecasts are frequently adjusted judgmentally, and these adjustments can help or harm depending on incentives, information

flows, and governance. Empirical and field-based work on forecasting in organizations shows that judgmental interventions are common and can become counterproductive when they systematically override statistical structure without accountability or feedback loops [8]. These findings highlight that forecasting quality is shaped not only by algorithms but also by how forecasts are produced, challenged, and used in decision-making routines.

Taken together, the literature frames forecasting as a system-critical capability: it affects operational decisions, buffers risk through inventory and capacity, and interacts with organizational behaviors and evaluation practices. Therefore, studies that compare forecasting approaches under controlled and consistent conditions contribute not only to methodological understanding but also to operational decision support, especially when the objective is to identify when improvements are feasible and when uncertainty is structurally irreducible. Introductory forecasting texts similarly emphasize that the central goal is “sensible use” of forecasting methods in context, rather than blind escalation of model complexity [9]. This perspective motivates empirical evaluations that focus on the decision relevance of accuracy and on the dependency of performance on data structure and operational constraints.

The increasing availability of machine learning tools has led to a rapid adoption of LSTM-based forecasting models in operational environments. However, when such models are applied without explicit assessment of data–model compatibility, substantial practical risks emerge. Deep learning models may converge numerically and produce seemingly smooth forecasts even when the underlying time series lacks a stable temporal structure. In such cases, the model effectively learns a regression-to-the-mean representation rather than meaningful signal dynamics.

From a decision-making perspective, this can create a false sense of reliability. Practitioners may interpret stable-looking forecasts as evidence of predictive power, while in reality, the model merely suppresses volatility. In inventory and procurement systems, such misaligned forecasts can lead to systematic underestimation of peaks, distorted safety stock calculations, increased stockout risk, or unnecessary capital tied up in excess inventory. Therefore, assessing whether data structure justifies the use of advanced learning-based models is not merely a methodological concern but a managerial necessity.

1.2. Overview of Forecasting Methodologies

Statistical forecasting methods form the classical foundation of time series analysis and are built on explicit probabilistic assumptions about how data evolve over time. These models typically decompose an observed series into components such as trend, seasonality, and random noise, allowing each element to be analyzed and modeled separately. Autoregressive and moving average models estimate future values as linear combinations of past observations and past errors, while integrated variants address non-stationarity through differencing [10]. Exponential smoothing methods assign decaying weights to historical data, making them effective for capturing slowly changing patterns with limited computational effort [11]. State space models further generalize these ideas by representing time series dynamics through latent variables and observation equations, enabling recursive estimation and uncertainty quantification [12]. Recent studies have reaffirmed the relevance of these approaches in modern applications, demonstrating their robustness in industrial forecasting and resource management tasks [13]. Extended seasonal models such as TBATS and advanced ETS variants have been shown to handle multiple seasonal patterns more effectively [14], while hybrid statistical formulations incorporating Kalman filtering have improved adaptability to evolving systems [15]. Comparative reviews published in recent years have emphasized that, despite advances in data-driven methods, classical statistical models remain competitive baselines due to their transparency and

stability [16]. However, their effectiveness depends strongly on stable data-generating mechanisms and may decline when relationships are highly nonlinear or subject to abrupt structural changes.

Heuristic forecasting methods are designed to provide practical and easily interpretable predictions without relying on strict statistical assumptions. These approaches typically use simple rules derived from empirical observations, such as moving averages, naïve persistence models, or manually adjusted forecasts based on domain expertise [17]. Their primary advantage lies in robustness and transparency, as the logic behind the prediction is straightforward and can be readily communicated to decision makers. Recent studies have shown that heuristic models remain valuable as benchmark references, particularly for evaluating more complex forecasting systems [18]. Heuristic methods are often applied in situations where data quality is poor, historical depth is limited, or rapid operational decisions are required [19]. However, their simplicity restricts their ability to capture complex temporal structures, interactions, or long-term dependencies, which limits their effectiveness for long-horizon or highly dynamic forecasting tasks [20].

Artificial intelligence-based forecasting methods range from shallow learning models to deep neural architectures. Shallow learning approaches, such as feedforward neural networks trained via backpropagation, rely on limited network depth and externally defined lag structures. In contrast, deep learning models introduce hierarchical representations and recurrent memory mechanisms that allow the direct modeling of sequential dependencies. In recent forecasting research, deep learning architectures have become dominant, reflecting the growing emphasis on modeling complex temporal dynamics.

Artificial intelligence-based forecasting methods aim to learn predictive relationships directly from data through adaptive algorithms. Techniques such as feedforward neural networks, support vector machines, and recurrent architectures, including LSTM models, can model nonlinear patterns and temporal dependencies without predefined functional forms [21,22]. Recent comparative studies have shown that recurrent and hybrid deep learning architectures often outperform classical statistical models on complex financial and industrial time series, particularly when nonlinear dynamics dominate [23]. Transformer-based and LSTM–Transformer hybrid models have further demonstrated improved robustness and long-horizon accuracy on large-scale forecasting benchmarks [24]. Empirical applications in demand forecasting and supply chain management have confirmed that AI-driven models can achieve higher predictive accuracy than traditional baselines when sufficient data are available [25]. At the same time, these methods introduce challenges related to limited interpretability, sensitivity to noise, and strong dependence on training data quality [26]. As a result, artificial intelligence models are often most effective when combined with careful preprocessing and complementary baseline methods to ensure stability and reliable evaluation [27].

1.3. Research Gaps in Practical Forecasting Applications

Based on the reviewed methods and practical considerations, several research gaps can be identified. First, despite the widespread use of simple forecasting techniques such as moving averages, exponential smoothing, and naïve persistence in procurement practice, limited empirical evidence exists on how much forecast accuracy is gained when these methods are replaced by LSTM-based models under identical data and evaluation settings. Second, there is insufficient analysis of whether the added complexity of LSTM models translates into consistent accuracy improvements for weekly operational data that are typically handled in spreadsheet-based workflows. Third, the behavior of LSTM forecasts relative to simple baselines remains underexplored, particularly in cases where demand patterns are noisy or weakly structured. Addressing these gaps is essential

for understanding when advanced forecasting models provide tangible benefits over established practical approaches.

Beyond methodological comparison, insufficient attention has been given to the operational consequences of misapplying complex forecasting models. While LSTM architectures are often promoted as universally superior, empirical evidence rarely examines how performance deteriorates when data lack sufficient structure. Without such understanding, organizations may adopt computationally intensive models that increase system complexity without improving decision quality.

To synthesize the key insights from the reviewed literature, Table 1 summarizes the role of forecasting in decision-making systems, the main categories of forecasting methodologies, and the identified research gaps. This structured overview highlights how data characteristics, rather than model complexity alone, determine forecasting performance and motivates the empirical analysis presented in the subsequent sections.

Table 1. Summary of the Theoretical Background.

Theme	Key Concept	Description	Main Implication
Role of Forecasting	Decision support	Forecasts directly influence production, procurement, inventory, and capacity decisions	Forecasting value should be evaluated by its impact on decisions, not only statistical accuracy
	Bullwhip effect	Forecast errors amplify upstream in supply chains	Inaccurate forecasts lead to systemic cost increases
	Inventory uncertainty	Forecast errors affect safety stock and service levels	Ignoring forecast error structure leads to biased inventory decisions
Forecasting Methods	Classical statistical models	AR, MA, ARIMA, exponential smoothing, state-space models	Transparent and stable, but sensitive to nonlinearity and structural changes
	Heuristic approaches	Naïve forecasts, moving averages, judgmental adjustments	Robust and interpretable baselines, but limited expressive power
	AI-based models	Neural networks, LSTM, hybrid deep learning models	Capture nonlinear and complex temporal patterns but suffer from limited interpretability
Evaluation Perspective	Data characteristics	Periodicity, noise level, and volatility determine forecastability	Data structure often dominates model complexity in determining accuracy
	Validation design	Leakage-free validation and appropriate metrics	Weak evaluation protocols can produce misleading performance results
Organizational Aspect	Judgmental forecasting	Human interventions in statistical forecasts	Can degrade accuracy if not governed by feedback and accountability
Research Gap	LSTM vs. simple methods	Limited evidence under identical data and evaluation settings	Performance gains of complex models are conditional rather than universal
	Weekly operational data	Spreadsheet-based planning environments are underrepresented in research	Practical relevance of advanced models remains uncertain
	Weakly structured time series	Highly noisy and irregular demand patterns	LSTM models may fail to outperform classical methods in such cases

This study contributes to the literature in several ways. First, it provides a controlled, structure-dependent comparison of classical forecasting methods (Naïve, MA, SES, Holt, ARIMA) and a uniformly configured LSTM on the same five real weekly time series under identical training and evaluation conditions, ensuring that performance differences are attributable to data characteristics rather than experimental design. Second, it demonstrates that forecasting accuracy is driven primarily by periodicity, volatility, and noise level rather than model complexity alone, consistent with bias–variance and signal-to-noise considerations. Third, it introduces a deliberately adapted LSTM framework for low-data environments, incorporating a causal moving-average input, omission of dropout due to constrained model capacity, and batch size of one to introduce implicit regularization

through stochastic gradient updates; these design decisions were motivated by stability and generalization requirements rather than arbitrary modification. Finally, the study validates the operational feasibility of LSTM-based forecasting in a real corporate weekly procurement context, emphasizing practical applicability over architectural novelty.

To respond to these issues, the remainder of the article is structured as follows. Section 2 introduces the forecasting problem and describes the underlying data, including their temporal resolution and basic characteristics. Section 3 presents the simple forecasting methods commonly applied in practice and evaluates their forecast accuracy on the given data. Section 4 introduces the adapted LSTM model and applies it to the same database under consistent experimental conditions. Finally, Section 5 compares the results across all methods, with particular emphasis on differences in forecast accuracy between the simple baselines and the LSTM approach.

In addition to methodological comparison, this study aims to clarify the boundary conditions under which LSTM models provide operational value. By explicitly linking data structure, forecast behavior, and decision relevance, the analysis contributes to a more informed and risk-aware application of advanced forecasting methods in practical planning environments.

2. Dataset Collection and Commonly Used Statistical Forecasting Methods

In this study, forecasting is applied to support planning and decision-making in a supply chain context, where reliable demand and material consumption estimates are essential for production, inventory, and procurement activities. To enable this analysis, historical operational data must first be collected and structured in a form suitable for both traditional and learning-based forecasting methods.

2.1. Dataset Description and Forecasting Context

The dataset used in this research was queried from the SAP (SAP S/4HANA 2023 FPS03) database of a real manufacturing company, with all sensitive product identifiers removed to protect confidentiality. It contains weekly demand histories for five distinct materials (Items A, B, C, D, E). The main objective of analyzing this dataset was to determine how accurately a uniformly constructed LSTM model and other statistical models can estimate future values across time series with differing structural properties and noise levels.

To enable both spreadsheet-based training and consistent benchmarking, the first 100 weeks of data were used for model training, and the subsequent 10 weeks served as the evaluation period. Although 10 weeks may appear short in absolute terms, such a horizon is meaningful in operational planning contexts where weekly forecasts directly influence decision-making.

In practical supply chain and procurement settings, typical forecast accuracy expectations vary by time horizon and application. For finished goods forecasting, realistic accuracy ranges are [5]:

- 1 week: 90–98% (used for operational planning and daily production)
- 1 month: 80–90% (short-term production and inventory planning)
- 3 months: 65–80% (capacity planning and tactical scheduling)
- 6 months: 55–70% (Sales & Operations Planning level)
- 1 year: 40–60% (strategic direction rather than detailed prediction)

For raw material procurement and supply planning, typical accuracy ranges tend to be higher due to aggregated order patterns and lower volatility [5]:

- 1 week: 95–99% (driven by specific production requirements)
- 1 month: 85–95% (order quantities and replenishment planning)

- 3 months: 75–90% (framework agreements/medium-term commitments)
- 6 months: 65–80% (capacity reservations)
- 1 year: 55–70% (volume forecasting at category level, not individually)

Accuracy is influenced by many factors. Elements that typically degrade performance include promotional activities, seasonal peaks, new product introductions, customer campaigns, and long supplier lead times. Conversely, forecast quality improves with stable demand, fixed ordering agreements, product family or category level aggregation, and rolling forecast practices.

Figure 1 illustrates the raw weekly time series of Items A through E and already reveals substantial differences in their behavior. Item A exhibits a highly regular and repeating pattern with low noise, making it visually easy to predict. Item B also shows a clearly recurring structure, although it contains intermittent zero values and sharp peaks. Despite these discontinuities, its overall behavior remains deterministic and appears well suited for forecasting.

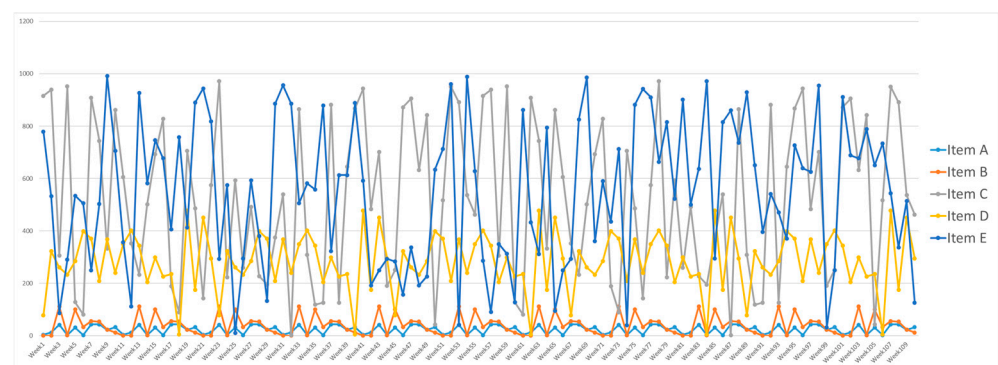


Figure 1. Raw weekly time series of items A–E.

Item C is characterized by extreme variability and irregular spikes, indicating a high noise level and weak structural regularity. Item D shows moderate variability with only occasional repetition, suggesting limited and unstable predictability. Item E is dominated by strong and erratic fluctuations with no visible repeating structure, and therefore represents a highly noise-driven and poorly predictable series.

Differences in absolute magnitudes across items should not influence the interpretation of their temporal behavior. For this reason, Figure 2 presents a normalized representation of all series. Each item was scaled relative to its own maximum value, while the minimum was consistently set to zero, even in cases where the observed minimum was higher. This choice reflects a realistic production scenario, where a given item may have weeks with no production or usage at all. In addition, this normalization step was required for the LSTM training process, as learning-based models are sensitive to scale differences and require comparable value ranges to ensure stable and effective optimization.

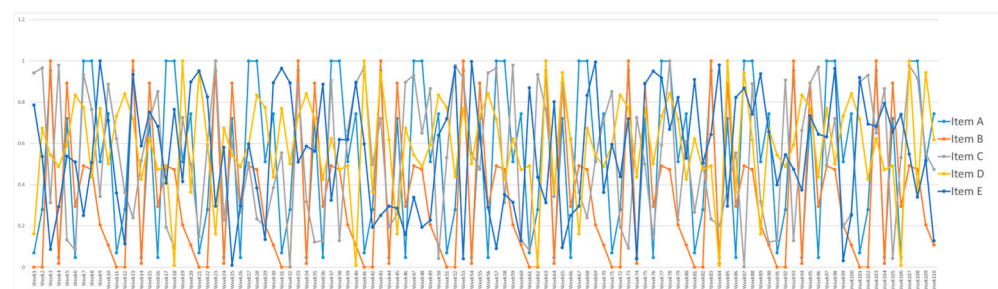


Figure 2. Normalized weekly time series of items A–E.

The statistical analyses of the data are presented in Table 2. Based on the regression results obtained from the Excel Data Analysis ToolPak (M365), the following observations may be made.

Table 2. Descriptive statistics and linear trend regression analysis of items A–E.

Item	Mean	Standard Dev.	Coefficient of Variation	Slope	R	R ²	Corrected R ²	Standard Error	Coefficient	t Value	p-Value
A	23.10	16.4314	0.7113	0.0261	0.0461	0.0021	−0.0081	16.4974	0.0261	0.4567	0.6489
B	39.00	38.8392	0.9959	0.0040	0.0030	0.0000	−0.0102	39.0367	0.0040	0.0293	0.9767
C	505.45	304.5492	0.6025	−0.6939	0.0661	0.0044	−0.0058	305.4297	−0.6939	−0.6558	0.5135
D	284.02	109.7698	0.3865	0.1521	0.0402	0.0016	−0.0086	110.2393	0.1521	0.3983	0.6913
E	544.37	281.5310	0.5172	0.7585	0.0782	0.0061	−0.0040	282.0979	0.7585	0.7762	0.4395

Across all five items, substantial differences in scale and variability are observed. The coefficient of variation indicates moderate to high relative volatility for each series. Items A and B exhibit particularly high dispersion relative to their means (CV = 0.71 and 0.99, respectively), while Items C, D, and E show moderate variability (CV between 0.39 and 0.60). This confirms that the datasets are characterized by non-negligible dispersion, which may influence forecasting performance.

With respect to linear trend diagnostics, it is observed that none of the estimated slope coefficients are statistically significant at the conventional 5% level. For all items, the *p*-values associated with the slope parameter substantially exceed 0.05 (ranging from 0.4395 to 0.9767). Furthermore, the 95% confidence intervals for all slope estimates include zero, indicating that the presence of a deterministic linear trend cannot be supported statistically. Although positive slopes are estimated for Items A, B, D, and E and a negative slope for Item C, these effects are small in magnitude and statistically indistinguishable from zero.

The explanatory power of the linear models is negligible. The R² values are extremely low for all items (between 0.0000 and 0.0061), and the adjusted R² values are negative in each case. This implies that the linear time index explains virtually none of the variance in the observed demand. The correlation coefficients (R) are also close to zero, further confirming the absence of a meaningful linear association between time and demand.

Taken together, these results indicate that the examined time series do not exhibit statistically significant linear trends over the analyzed horizon. Instead, variability appears to be driven primarily by irregular fluctuations rather than systematic upward or downward movement. Consequently, the forecasting behavior observed in subsequent analyses cannot be attributed to strong deterministic trend components but must instead be interpreted in light of volatility, periodicity, and structural characteristics of the individual series. The Excel Data Analysis ToolPak was used, and the following observations may be made.

2.2. Commonly Used Statistical Forecasting Methods

In operational forecasting practice, a small set of simple methods continues to dominate daily decision-making, despite the availability of more advanced techniques. These approaches are still widely applied because they are easy to implement, well understood, and deliver reliable performance across a broad range of use cases. Methods such as moving averages and exponential smoothing, including Holt's formulation, have remained in use for decades, as they provide stable and consistent forecasts regardless of the specific operational context. Their robustness and transparency make them trusted tools in practice and natural reference points for evaluating more complex forecasting models [2,11].

The naïve persistence method assumes that the future value of a time series equals its most recent observed value. Formally, the forecast at time $t + 1$ is given by:

$$\hat{y}_{t+1} = y_t \quad (1)$$

This method contains no parameters and no learning mechanism. It is widely used because it represents the simplest possible forecasting rule and often reflects implicit decision-making in practice. Its importance lies in its role as a benchmark, since any advanced model is expected to outperform it in forecast accuracy to justify additional complexity [2].

The moving average method estimates future demand as the arithmetic means of the most recent n observations. The forecast is defined as

$$\hat{y}_{t+1} = \frac{1}{n} \sum_{i=0}^{n-1} y_{t-i} \quad (2)$$

This approach smooths short-term fluctuations and reduces random noise. It is frequently used in operational planning because it is easy to compute, transparent, and effective when demand exhibits no strong trend or seasonality.

The simple exponential smoothing (SES) method improves upon moving averages by assigning exponentially decreasing weights to past observations [11]. The forecast is updated recursively as

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha) \hat{y}_t \quad (3)$$

where $\alpha \in (0, 1)$ is the smoothing parameter. Recent observations receive higher importance, making the method more responsive to changes. SES is commonly applied to stable time series without pronounced trends and is widely supported in spreadsheet-based tools.

The Holt exponential smoothing method extends SES by explicitly modeling a trend component [11]. It is defined by

$$\begin{aligned} l_t &= \alpha y_t + (1 - \alpha)(l_{t-1} + b_{t-1}) \\ b_t &= \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1} \\ \hat{y}_{t+h} &= l_t + hb_t \end{aligned} \quad (4)$$

where l_t denotes the level, b_t the trend, $\alpha, \beta \in (0, 1)$ are smoothing parameters and h denotes the forecast horizon and represents the number of time steps ahead for which the prediction is made. Holt's method is used when demand shows a consistent upward or downward movement. It remains popular in practice because it balances simplicity with the ability to capture systematic trends.

Together, these methods represent the most widely adopted forecasting techniques in operational environments. They are used not because they are optimal in all cases, but because they provide stable, interpretable, and computationally efficient baselines against which more complex models, such as LSTM networks, can be evaluated.

2.3. ARIMA Forecasting Method

In addition to classical statistical forecasting techniques, a widely established statistical method was applied to support data analysis and to provide a structured baseline for comparison. Among these methods, the Autoregressive Integrated Moving Average model, commonly referred to as ARIMA, was selected due to its long-standing theoretical foundation and its extensive use in time series forecasting. ARIMA models have been thoroughly studied and continue to serve as a standard reference in both academic research and practical applications [2].

The ARIMA framework describes a univariate time series through three components. The autoregressive component captures linear dependence between current and past observations. The moving average component models the influence of past forecast errors on current values. The integrated component addresses nonstationarity by applying

differencing to the original series. Together, these elements enable ARIMA models to represent a wide range of temporal behaviors under relatively simple assumptions.

An ARIMA model is denoted by the parameter triplet (p, d, q) , whose interpretation is summarized in Table 3. Here, p represents the order of the autoregressive component, d denotes the degree of differencing applied to achieve stationarity, and q specifies the order of the moving average component. Model identification typically begins with an examination of the time series properties, including trend presence, variance stability, and autocorrelation structure. Stationarity is a critical requirement for ARIMA modeling, and differencing is applied when trends or slow level shifts are detected.

Table 3. Interpretation of ARIMA Model Parameters.

P.	Typical Values	Meaning
p	0, 1, 2, 3	Order of the autoregressive component. It defines how many past observations are used to explain the current value. Higher values indicate stronger or longer linear dependence on previous time steps.
d	0, 1, 2	Degree of differencing applied to the time series to achieve stationarity. A value of zero indicates that the series is already stationary, while higher values remove trends or slow level shifts.
q	0, 1, 2, 3	Order of the moving average component. It represents how many past forecast errors influence the current observation, capturing short-term shock effects.

In this study, ARIMA models were used primarily for exploratory data analysis and as a classical forecasting benchmark. Before model estimation, each time series was inspected visually and through summary statistics to assess its structural characteristics. Stationarity was formally evaluated using the Augmented Dickey–Fuller (ADF) test at a 5% significance level. If non-stationarity was detected, first- or second-order differencing was applied ($d \in \{0, 1, 2\}$). The autoregressive and moving average orders were determined through a grid search over $p \in \{0, 1, 2, 3\}$ and $q \in \{0, 1, 2, 3\}$, resulting in up to 48 candidate ARIMA (p, d, q) specifications per series. Parameters were estimated using maximum likelihood, and the Akaike Information Criterion (AIC) was applied to select the optimal specification, ensuring a balance between goodness-of-fit and model parsimony. Only convergent models satisfying stationarity and invertibility conditions were retained.

Residual diagnostics were performed on the selected models to verify adequacy. The Ljung–Box test (lag = 10) was applied to confirm the absence of significant residual autocorrelation, and residual ACF plots were visually inspected to ensure that remaining linear dependence was minimal. Forecasts were generated analytically for a 10-week horizon using the fitted ARIMA models without recursive re-estimation. The same chronological train–test split and evaluation protocol was applied as in the LSTM experiments to ensure a fair and transparent benchmark comparison.

Within this study, ARIMA modeling served two main purposes. First, it provided an interpretable statistical description of the temporal dynamics of each item, allowing differences in stability and persistence to be identified through model order and fit quality. Series with regular behavior were typically well represented by low-order ARIMA models, while highly volatile items showed weaker performance.

Second, ARIMA forecasts were used as a conservative reference baseline. Due to their linear assumptions and limited flexibility, ARIMA models offer a meaningful benchmark against which the performance of more complex forecasting approaches can be evaluated. Performance gains beyond this baseline can therefore be attributed to modeling capabilities not captured by classical linear structure.

Overall, ARIMA models were found to perform reliably on stable series and short horizons, while their limitations became apparent for irregular and highly noisy data. Nevertheless, their transparency and well-understood properties make them a valuable analytical tool and an essential point of comparison in forecasting studies.

3. LSTM-Based Forecasting and Methodological Adaptations

In this section, the methodology applied for LSTM-based time series forecasting is described. The objective was to construct a unified modeling framework that can be applied consistently across multiple items and temporal resolutions, while enabling multi-step prediction and clear interpretation of model behavior. A Long Short-Term Memory network was selected due to its ability to capture temporal dependencies in sequential data and to retain relevant information over longer horizons [23,24].

The analysis was conducted on multiple weekly time series for five items denoted as A through E. All series were preprocessed individually to ensure numerical stability and comparability across items. Each time series was normalized to the interval $[0, 1]$ using a normalization function, where the minimum value was fixed at zero, and the maximum was defined by the observed peak value. This normalization reflects realistic production conditions, where zero usage may occur at any time, and was also required to ensure stable and efficient LSTM training, particularly for items with large amplitude variation.

Although a standard LSTM architecture was used as the foundation, several deliberate methodological adaptations were introduced in order to align the model with the properties of the available data and the requirements of the forecasting task. These adaptations should not be interpreted as minor implementation details, but as explicit modeling and training decisions motivated by limited data availability, heterogeneous noise characteristics, and the need for stable prediction [23].

A single-layer LSTM architecture with 32 memory units was applied (see Figure 3), followed by a fully connected output layer whose dimensionality matched the selected forecast horizon. The objective was not to maximize model complexity, but to evaluate whether a compact and uniformly specified architecture can reliably learn different temporal structures across items with varying levels of noise and regularity.

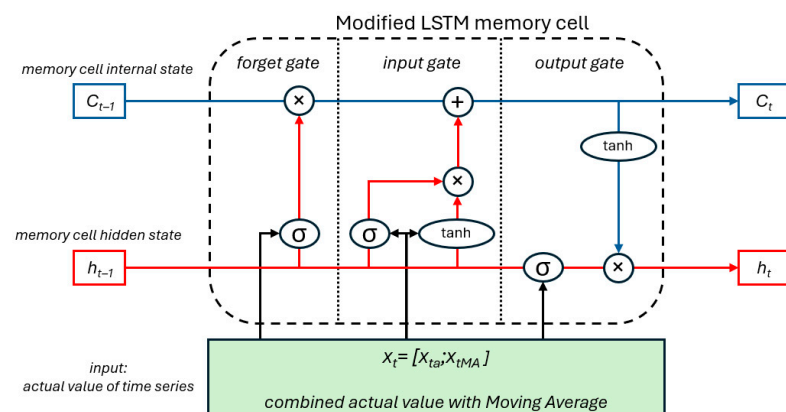


Figure 3. Modified LSTM model for improved trend approximation.

For the forecasting process, training samples were generated using a sliding window approach. In the weekly setting, a window of 12 past observations was used to predict the subsequent 12 weeks. This configuration was selected after repeated experimental calibration, as it consistently provided the most stable validation performance across the examined items. Notably, this result was somewhat unexpected, since the moving-average optimization did not identify 12 weeks as an optimal smoothing horizon for any item, indicating that the effective memory length of the LSTM does not necessarily coincide with

classical statistical parameters. It should also be emphasized that the optimal window size is data-dependent and may differ for other datasets.

Due to the sequential nature of time series data, all samples were split strictly in chronological order rather than randomly. The first 60 percent of the observations were used for training, the subsequent 20 percent for validation, and the remaining 20 percent for testing. This separation ensured that model evaluation reflected realistic forecasting conditions and prevented information leakage from future observations.

Modifications and Adaptations of LSTM Method

Although this configuration follows a standard LSTM modeling framework, several deliberate methodological adaptations were introduced to better align the model with the characteristics of the data and the requirements of the forecasting task.

The first and most substantial modification is the introduction of a moving average representation as an additional input feature alongside the raw time series as a second dimensional input ($X_t = [X_{ta}; X_{tMA}]$). This modification represents a data representation strategy rather than a change to the LSTM cell itself. The moving average was computed using a strictly causal rolling window, ensuring that only past information was used and preventing any form of information leakage. By explicitly providing a smoothed trend component, the network was relieved from learning basic low frequency structure directly from noisy observations. This is particularly important in small datasets with large outliers, where jointly learning trend and volatility can lead to unstable parameter updates and poor generalization.

Second, the Dropout layer was intentionally omitted from the network. In recurrent architecture, dropout acts as an explicit regularization mechanism by randomly suppressing activations during training, which can be beneficial in large or highly expressive models. In the present setting, however, the model capacity was intentionally kept low and no systematic divergence between training and validation error was observed under time-ordered validation. Under these conditions, additional regularization would unnecessarily constrain the representational capacity of the network. Preserving full information flow through the LSTM cells was therefore preferred, allowing the model to converge more accurately to subtle temporal dependencies present in structured items.

Third, a batch size of one was applied during training, deviating from common mini batch conventions. This choice results in highly noisy gradient updates, as each parameter update is based on a single sequence. From an optimization perspective, this stochasticity introduces implicit regularization, which can improve generalization when learning from small datasets. At the same time, such training dynamics can be unstable, particularly in recurrent networks. To ensure convergence, careful learning rate selection and early stopping were employed. This configuration allowed the model to explore flatter regions of the loss landscape without relying on explicit regularization layers.

Finally, a direct multi-step forecasting strategy was adopted, in which the output layer generates the full forecast horizon in a single forward pass. This approach differs fundamentally from recursive one step forecasting, where predictions are iteratively fed back into the model. Recursive strategies are known to suffer from error accumulation, as small inaccuracies propagate and amplify over time. By predicting all future steps simultaneously, this source of drift was reduced. While direct multi output forecasting introduces the challenge of jointly optimizing multiple horizons within a single loss function, it provides a more controlled and interpretable evaluation of long-range predictive behavior.

Taken together, these methodological adaptations form a coherent and purpose-driven LSTM framework tailored to the structure and constraints of the studied data. They directly

influence learning dynamics, stability, and forecast behavior, and they provide a transparent basis for the comparative analysis presented in the subsequent sections.

4. Training the LSTM Model

In this chapter, the training procedure of the LSTM model is described in detail. The objective of the training phase was to ensure stable convergence while allowing the model to fully exploit its capacity to learn temporal patterns from the available data. Emphasis was placed on robustness and generalization rather than architectural complexity.

For this study, a single-layer LSTM architecture with 32 memory units was applied, followed by a fully connected Dense output layer whose dimensionality matched the selected forecast horizon. This compact architecture was chosen deliberately, as the goal was not to maximize depth or parameter count, but to evaluate how effectively a simple and uniform LSTM configuration can learn different temporal structures across items. The model was optimized using the Adam optimizer (with a 0.001 learning rate, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$) and trained with mean squared error as the loss function, which is well suited for continuous forecasting tasks.

Training was performed using an early stopping strategy in order to prevent unnecessary epochs and reduce the risk of overfitting. Validation loss was monitored throughout training, and training was terminated when no improvement was observed for 20 consecutive epochs. The model parameters corresponding to the lowest validation error were automatically restored. A maximum of 800 epochs was allowed, although early stopping typically resulted in significantly shorter effective training durations.

A batch size of one was used during training. This choice ensured that parameter updates were performed after each individual sequence, preserving the temporal structure of the data and introducing beneficial stochasticity into the optimization process. Model performance during training was evaluated using a chronologically separated validation set, ensuring that only past information was used to predict future observations.

In the following subsections, the trained model is evaluated separately for each item. Results are presented with particular focus on learning dynamics, forecast behavior, and error characteristics. Item A serves as an initial reference case, as it represents a highly structured and minimally noisy time series. For this item, training curves converge rapidly, and both weekly and monthly forecasts closely follow the observed values. The model achieved R^2 values close to 1.00, indicating very high explanatory power, and the prediction errors remain small and uniformly distributed across the forecast horizon. This behavior illustrates an ideal scenario in which the LSTM model can fully exploit repeating temporal patterns and produce stable forecasts.

4.1. Item A and B Forecasting: An Easily Learnable Time Series

The visual results further confirm the strong learning behavior observed for Item A. As shown in Figure 4, the loss curves display a smooth and monotonic decrease for both training and validation error, with the two curves closely aligned throughout the training process. This indicates stable convergence and the absence of overfitting, as no divergence between training and validation loss is observed. The rapid reduction in mean squared error in the early epochs suggests that the dominant temporal structure is learned efficiently, while the gradual flattening at later epochs reflects fine scale adjustment rather than unstable optimization.

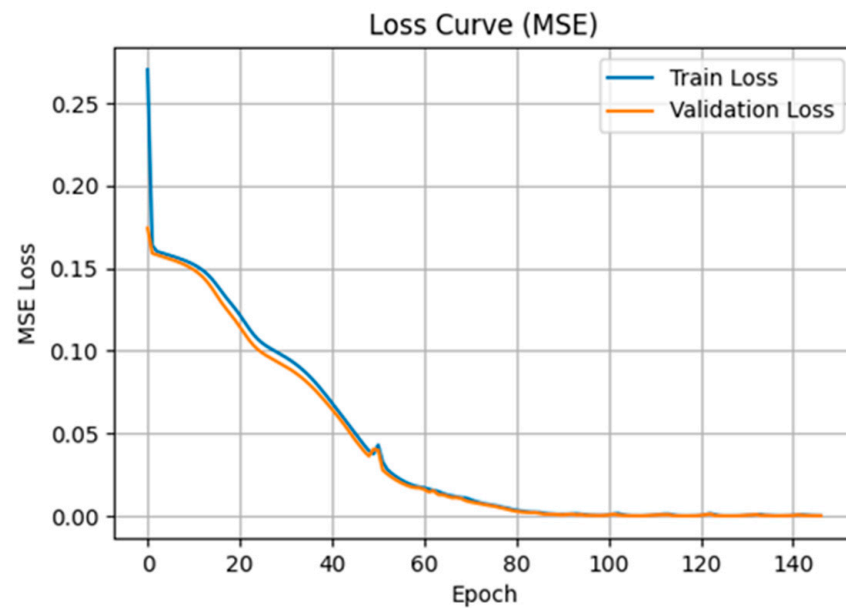


Figure 4. Training phase on weekly data on Item A.

The prediction error heatmap in Figure 5, provides a compact and informative representation of forecast errors across both time and horizon steps. Each row corresponds to a test sequence, while each column represents a forecast step within the horizon. For Item A, the heatmap is dominated by near zero values with alternating but balanced positive and negative errors. This pattern indicates that errors are small, unbiased, and consistent across horizons. The absence of systematic drift or horizon dependent degradation demonstrates that the direct strategy is effective for this structured series.

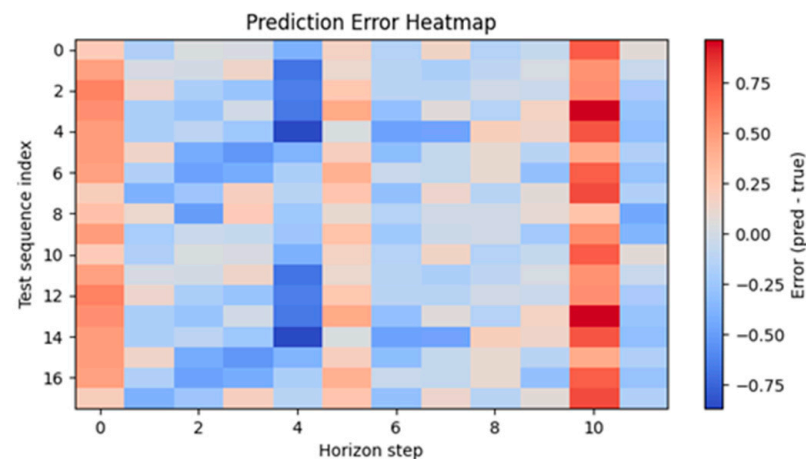


Figure 5. Heatmap on Item A.

The direct comparison of weekly observations and forecasts of item quantity shown in Figure 6, illustrates that the predicted values closely follow the cyclical pattern of the original series. The forecasted segment continues the established periodic structure without phase shift or amplitude distortion. This alignment confirms that the model has learned the underlying generative pattern rather than merely extrapolating recent values. Together, these visualizations provide consistent evidence that Item A represents an ideal scenario for LSTM-based forecasting.

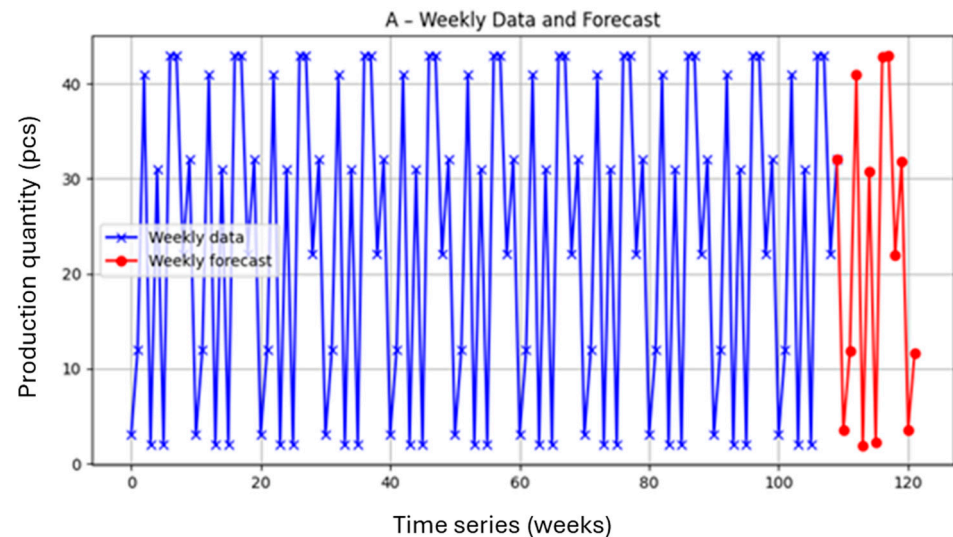


Figure 6. 11-week quantity forecast on Item A.

Item A represents an ideal case for LSTM-based forecasting, as the time series exhibits a stable and near deterministic repeating structure. Under such conditions, temporal dependencies are consistently preserved, allowing the model to learn and reproduce the underlying pattern with high accuracy. It is observed that reliable forecasts can be achieved even when only a limited portion of the historical data is available for training. This result confirms that, when strong regularity and low noise are present, the LSTM model is able to generalize effectively and capture the essential dynamics of the series.

Item B also exhibits a recurring temporal structure, although it is characterized by frequent alternation between zero values and sharp peaks. Despite these abrupt transitions, the repetition pattern remains sufficiently consistent for the model to learn. The results show that peak amplitudes are approximated with high accuracy, indicating that the dominant signal strength is well captured. Forecasts on the monthly scale closely approximate the observed values, while weekly predictions occasionally display a slight phase shift relative to the observed series. Throughout training, convergence remains stable and no signs of overfitting are observed. Figure 7 confirms that the model can handle discontinuous yet deterministic patterns effectively.

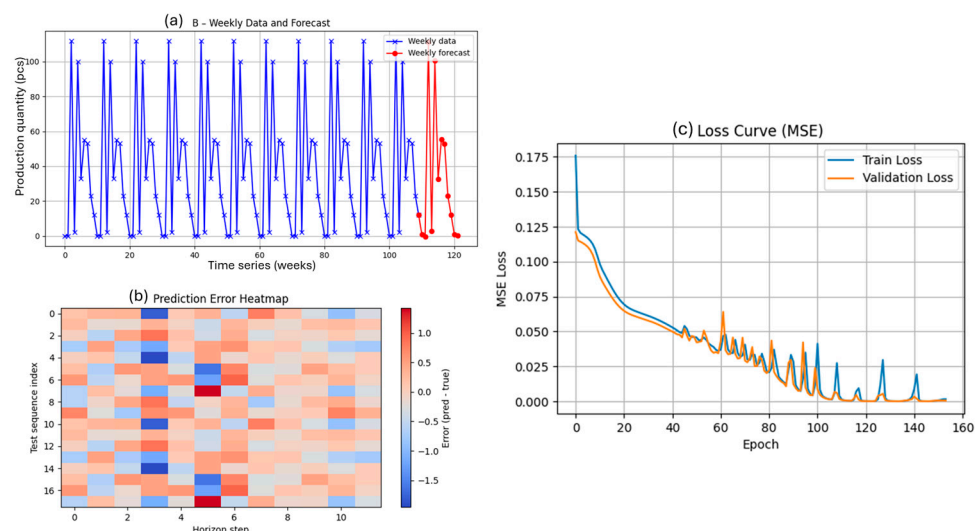


Figure 7. Weekly quantity forecast (a) heatmap (b) and training phase (c) on Item B.

4.2. Item C, D and E Forecasting: Low Prediction and Low Structured Time Series

Item C represents a high amplitude and weakly structured time series, with weekly values ranging from very small magnitudes to extreme peaks close to one thousand. Such behavior introduces severe instability and strong outliers, which significantly complicate the learning process. As shown in Figure 8, the model responds by applying pronounced smoothing, resulting in forecasts that capture only the general direction of the series while failing to reconstruct individual peaks. Even after temporal aggregation, performance remains poor, as the underlying irregularity persists despite reduced noise. This behavior is reflected in consistently negative R^2 values, indicating that predictions are inferior to simple baseline estimates. Overall, no stable or repeatable pattern can be identified as to allow reliable learning, and the results demonstrate that the model is not suitable for forecasting data dominated by extreme variability and weak structural regularity.

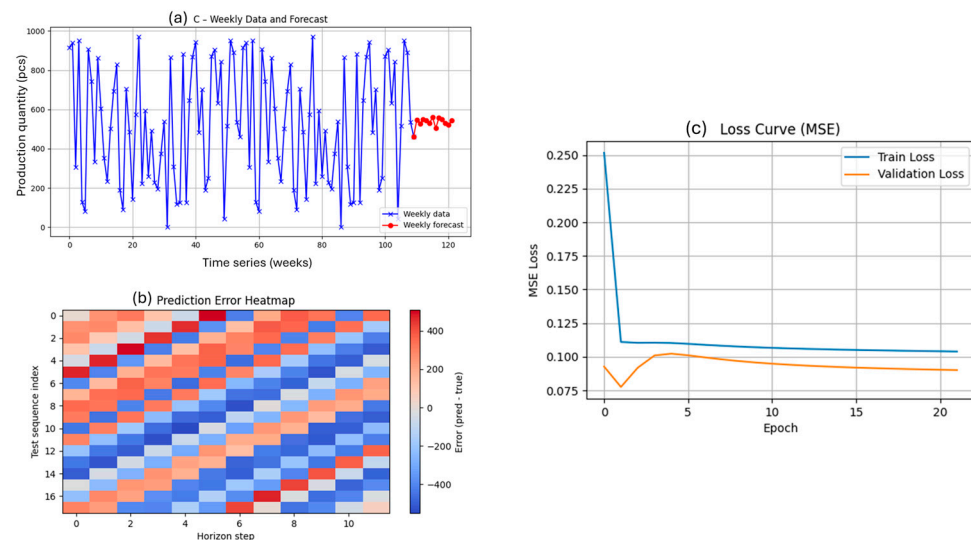


Figure 8. Weekly quantity forecast (a), heatmap (b) and training phase (c) on Item C.

Item D exhibits a moderate level of noise combined with occasional periodic behavior. While the series is less erratic than Item C, the underlying structure remains weak and inconsistent, which limits the ability of a generic LSTM model to generalize effectively. As illustrated in Figure 9, the results show systematic behavior, as peaks are consistently underestimated while troughs tend to be overestimated. This indicates a regression toward the mean rather than precise pattern reproduction. Forecasts at the monthly level appear excessively smoothed, suggesting that aggregation further suppresses already weak structural signals. The coefficient of determination fluctuates around zero, confirming that predictive performance remains close to that of a naïve baseline. Overall, the model achieves only limited accuracy for Item D, as the temporal pattern is not sufficiently strong or stable to support a reliable learned representation.

Item E represents the most challenging case among all analyzed series due to its extremely erratic and highly volatile behavior. Weekly values fluctuate roughly between 300 and 1400, while monthly aggregates reach ranges of approximately 800 to 3800, leaving little consistent structure for the model to exploit. As shown in Figure 10, the model fails to reproduce extreme values and instead collapses toward overly smoothed estimates. Prediction errors remain large and unstable, and forecast quality deteriorates further as peaks dominate the series. Monthly level forecasts appear even more distorted, as aggregation amplifies the influence of extreme observations rather than stabilizing the pattern. This behavior is reflected in drastically negative R^2 values, reaching as low as minus seven, indicating performance far below simple baseline expectations. Overall, Item E

demonstrates that when a time series is dominated by irregular spikes and lacks repeatable structure, reliable forecasting is not achievable with the applied modeling approach.

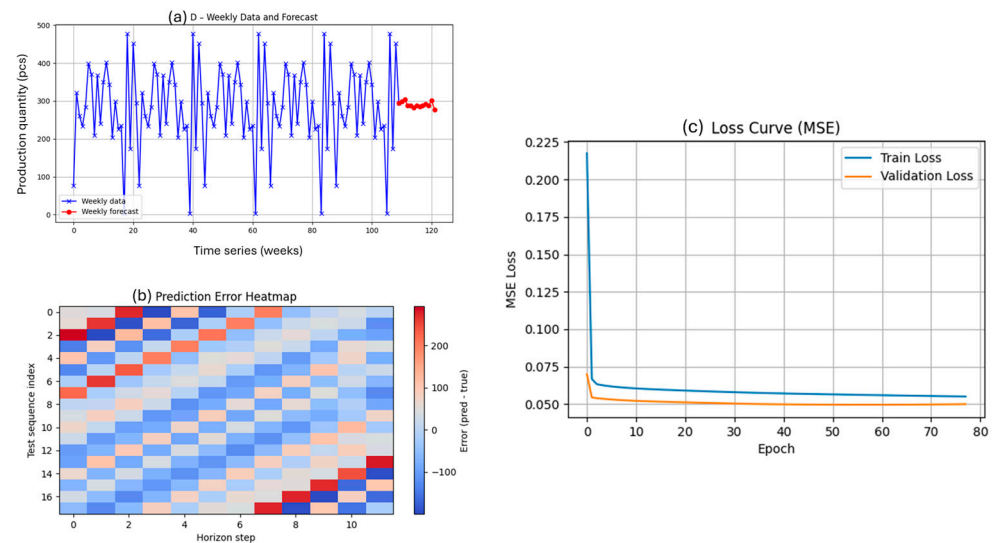


Figure 9. Weekly quantity forecast (a), heatmap (b), and training phase (c) on Item D.

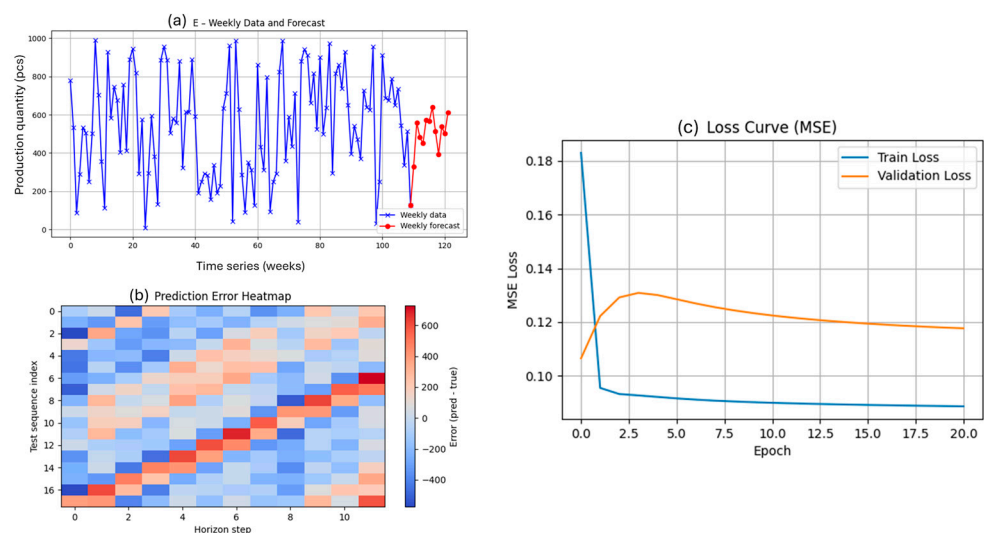


Figure 10. Weekly quantity forecast (a), heatmap (b), and training phase (c) on Item E.

This chapter shows that the performance of the applied LSTM model is strongly driven by the internal characteristics of the time series, including noise intensity, regularity, and the presence of stable temporal patterns. Although an identical model architecture and training procedure were used for all items, markedly different learning behaviors and outcomes were observed. Table 3 summarizes the training results across all items and provides a concise overview of how the LSTM model responds to varying data structures.

Table 4 reveals a clear contrast between structured and unstructured series. Items A and B achieve R^2 values approaching unity and very low error levels, although training times are longer due to the presence of strong and learnable temporal structure. In contrast, Items C and E converge quickly but yield poor predictive performance, as reflected by extremely high error values and negative R^2 scores. Item D shows intermediate behavior, with modest improvement over baseline and limited explanatory power. These results emphasize that training time alone does not indicate model success, and that meaningful learning occurs only when sufficient structure is present in the data.

Table 4. Training Summary of LSTM.

Item	MSE	R ²	Training Time (s)
A	0.07	1.00	32.83
B	0.12	1.00	32.31
C	112,356.71	−0.11	6.87
D	10,458.27	0.06	14.34
E	69,993.03	−0.20	11.65

All experiments were implemented in Python 3.12 using standard scientific computing libraries. Model development and training were carried out using TensorFlow 2.20.0. The experiments were executed in a CPU-based environment, and consistent results were ensured through fixed random seeds, with the specific hardware: AMD Ryzen 7 5700X (8-core, 4.65 GHz) CPU, 32 GB of DDR4 RAM, and an NVIDIA GeForce RTX 4060 Ti GPU (8 GB VRAM). To support reproducibility and transparency, the complete source code and the datasets used in this study are publicly available at: https://github.com/Strigula93/LSTM_PO_FC, accessed on 1 January 2026.

5. Comparative Analysis of Forecasting Models

In this chapter, the previously presented modified LSTM-based forecasting model is systematically compared with classical statistical forecasting methods, including ARIMA. The comparison is conducted under identical data and evaluation conditions in order to assess relative performance and to examine how model behavior differs across time series with varying structural properties.

In Table 5, the original observed values from week 101 to week 110 are presented for all items. These data points serve as the evaluation window and are used as the target values for forecasting with the alternative statistical methods. Forecast performance is assessed by examining how closely each method reproduces the observed values over this period.

Table 5. Actual weekly values for Items A–E (Weeks 101–110).

Item	W101	W102	W103	W104	W105	W106	W107	W108	W109	W110
A	3.00	12.00	41.00	2.00	31.00	2.00	43.00	43.00	22.00	32.00
B	0.00	0.00	112.00	2.00	100.00	33.00	55.00	53.00	23.00	12.00
C	872.00	905.00	632.00	842.00	42.00	517.00	950.00	891.00	536.00	462.00
D	343.00	204.00	298.00	226.00	235.00	4.00	478.00	174.00	451.00	295.00
E	911.00	688.00	677.00	788.00	650.00	734.00	543.00	337.00	514.00	126.00

The first method considered is the simplest baseline, the naïve forecasting approach, which is presented in Table 6. Using this forecasting method, the following predicted values were obtained.

Table 6. Naïve forecasting values for Items A–E (Weeks 101–110).

Item	W101	W102	W103	W104	W105	W106	W107	W108	W109	W110
A	32.00	32.00	32.00	32.00	32.00	32.00	32.00	32.00	32.00	32.00
B	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00
C	251.00	251.00	251.00	251.00	251.00	251.00	251.00	251.00	251.00	251.00
D	402.00	402.00	402.00	402.00	402.00	402.00	402.00	402.00	402.00	402.00
E	250.00	250.00	250.00	250.00	250.00	250.00	250.00	250.00	250.00	250.00

Table 7 presents the results obtained using the moving average method. The moving average window length was determined using Excel’s moving average function, which

selects the number (MA n.) of past observations that minimizes the overall deviation from the actual values. This procedure identifies the optimal averaging horizon by evaluating how different window sizes affect the total forecasting error relative to the original data.

Table 7. MA forecasting values for Items A–E (Weeks 101–110).

Item	W101	W102	W103	W104	W105	W106	W107	W108	W109	W110	MA n.
A	25.33	25.04	27.92	27.54	30.73	29.20	27.47	28.15	27.67	27.97	9
B	35.75	35.75	31.90	27.68	28.62	31.94	31.18	30.26	29.93	30.39	4
C	406.25	387.06	308.58	338.22	360.03	348.47	338.83	346.39	348.43	345.53	4
D	327.75	333.22	325.00	319.37	333.17	328.81	339.91	338.65	330.74	331.11	8
E	512.22	509.02	513.25	529.17	507.18	492.54	477.71	424.68	468.42	492.69	9

Table 8 presents the forecasting results obtained using simple exponential smoothing. The smoothing parameter alpha was optimized separately for each item using the built-in Excel Solver, with the objective of minimizing the overall forecasting error relative to the observed values.

Table 8. SES forecasting values for Items A–E (Weeks 101–110).

Item	W101	W102	W103	W104	W105	W106	W107	W108	W109	W110	Alpha
A	29.79	29.79	29.79	29.79	29.79	29.79	29.79	29.79	29.79	29.79	0.396002
B	26.23	26.23	26.23	26.23	26.23	26.23	26.23	26.23	26.23	26.23	0.466610
C	658.20	658.20	658.20	658.20	658.20	658.20	658.20	658.20	658.20	658.20	0.009606
D	271.79	271.79	271.79	271.79	271.79	271.79	271.79	271.79	271.79	271.79	0.025383
E	670.67	670.67	670.67	670.67	670.67	670.67	670.67	670.67	670.67	670.67	0.006336

Table 9 presents the forecasting results obtained using the Holt exponential smoothing method. Both the alpha and beta parameters were optimized using the Excel Solver in order to minimize forecasting error. As can be observed from the table and will be further confirmed by the graphical results, this method provides the strongest performance among the simple statistical forecasting approaches considered.

Table 9. Holt exponential smoothing method values for Items A–E (Weeks 101–110).

Item	W101	W102	W103	W104	W105	W106	W107	W108	W109	W110	Alpha	Beta
A	9.50	12.00	14.50	17.00	19.50	22.00	24.50	27.00	29.50	32.00	0.022644	0.656728
B	0.18	0.19	0.19	0.20	0.21	0.21	0.22	0.22	0.23	0.23	0.001787	0.978542
C	771.04	741.66	712.28	682.90	653.52	624.14	594.76	565.38	536.00	506.62	0.099346	0.999983
D	193.53	204.79	216.04	227.29	238.55	249.80	261.05	272.31	283.56	294.81	0.053509	0.269572
E	843.38	793.32	743.25	693.19	643.13	593.06	543.00	492.94	442.87	392.81	0.042374	0.341511

Table 10 presents the forecasting results generated using the ARIMA method, implemented through an existing Python (Python 3.14.3) library. The selected (p, d, q) parameter values for each item are reported at the end of the corresponding rows in the table.

Table 11 presents the forecasting results obtained using the previously described LSTM model, evaluated over the same weekly horizon as the statistical methods to ensure a consistent and fair comparison.

Table 10. ARIMA forecasting values for Items A–E (Weeks 101–110).

Item	W101	W102	W103	W104	W105	W106	W107	W108	W109	W110	<i>p</i>	<i>d</i>	<i>q</i>
A	8.45	18.62	21.82	22.83	23.15	23.25	23.28	23.29	23.29	23.29	1	1	2
B	49.90	23.23	52.03	24.65	52.54	25.18	52.58	25.46	52.48	25.65	2	0	2
C	600.17	534.94	444.99	474.00	493.93	481.87	476.66	479.44	479.25	477.22	2	2	2
D	252.71	297.48	285.00	289.41	288.84	289.74	290.20	290.80	291.35	291.92	1	2	2
E	547.57	568.72	570.90	571.79	572.58	573.37	574.15	574.94	575.73	576.51	1	2	2

Table 11. Modified LSTM forecasting values for Items A–E (Weeks 101–110).

Item	W101	W102	W103	W104	W105	W106	W107	W108	W109	W110
A	4.50	11.80	40.10	−0.40	30.20	4.30	45.00	44.40	22.70	31.00
B	2.40	−2.20	111.60	4.90	99.60	31.50	54.80	54.00	24.40	12.70
C	511.20	486.10	495.00	435.50	519.60	457.80	461.40	457.20	514.50	459.00
D	246.50	274.40	264.30	286.70	268.50	273.20	292.90	281.00	276.20	258.70
E	549.00	533.30	537.10	590.80	571.10	567.10	558.70	559.50	600.50	512.00

For comparison purposes, as previously described, the evaluation is based on the sum of deviations between the normalized actual values from Table 4 and the corresponding forecasts over the ten predicted weeks. This aggregated measure reflects how closely each method reproduces the observed behavior across items. Smaller values indicate better forecasting performance. The resulting comparison values are reported in Table 11.

Table 12 presents a compact and consistent overview of forecast performance across all five items based on the aggregated normalized error over the ten-week evaluation horizon, where smaller values indicate more accurate forecasts. This metric can be interpreted in practical terms by noting that the evaluation spans ten weeks. By scaling the aggregated error accordingly, an approximate percentage of the average weekly deviation can be obtained, allowing the results to be expressed in an intuitive and operationally meaningful form.

Table 12. Comparison of Forecasting Methods Across Items A–E.

Item	Naive	Moving A.	SES	Exp. HOLT	ARIMA	Mod. LSTM	Best Forecasting Method
A	3.5116	3.2916	3.4651	2.8256	3.0374	0.3070	Mod. LSTM
B	3.0179	2.8715	2.8214	3.4702	2.5368	0.1170	Mod. LSTM
C	4.6883	3.8655	2.3364	2.0039	2.7169	2.8878	Exp. HOLT
D	3.2678	2.3994	2.1381	2.0204	2.2324	2.2326	Exp. HOLT
E	3.7497	1.9676	1.6428	0.9845	1.8410	1.8267	Exp. HOLT

For instance, in the case of Item A, the LSTM model yields an aggregated error of 0.3070, which corresponds to an average weekly deviation of roughly 3 percent. This implies that the predicted values remain very close to the actual observations on a week-by-week basis. In contrast, the statistical methods produce aggregated errors in the range of 2 to 4 for the same item, corresponding to average weekly deviations of approximately 20 to 40 percent, which indicates a substantially lower level of precision.

For Items A and B, the modified LSTM model clearly outperforms all statistical alternatives by a wide margin. The markedly lower error values demonstrate that the strong periodic structure present in these series is effectively learned and reproduced by the model. Traditional methods, while internally consistent, are unable to achieve comparable

accuracy for these items. This performance difference is further illustrated in Figure 11 for Item A, where the alignment between predicted and observed values is visually apparent.

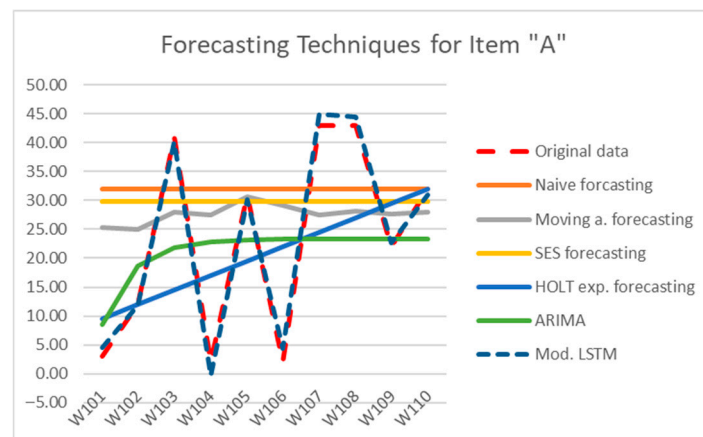


Figure 11. Comparison of forecasting techniques for Item A.

For Items C, D, and E, a different behavior is observed on Table 10 and Figure 12. These series are characterized by weaker structure and higher noise levels, and the modified LSTM does not achieve the lowest error. Instead, the Holt exponential smoothing method yields the best performance across all three items. This suggests that for irregular or moderately structured series, simpler statistical models with explicit trend handling can be more robust than learning-based approaches.

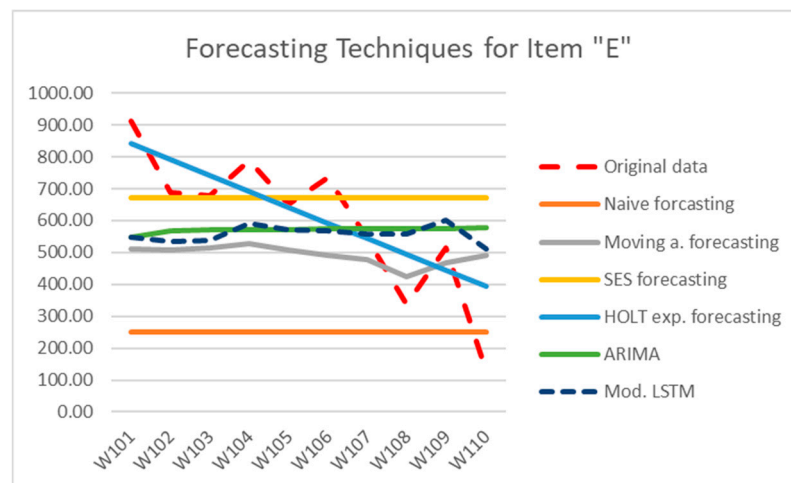


Figure 12. Comparison of forecasting techniques for Item E.

Across all items, the naive method consistently produces the highest error values, confirming its role as a baseline rather than a competitive forecasting solution. Moving average and simple exponential smoothing methods generally improve upon the naïve baseline but remain inferior to Holt or LSTM depending on the data structure.

ARIMA is the most complex statistical method included in the comparison, yet its performance does not consistently exceed that of simpler approaches and is often matched or surpassed by Holt exponential smoothing. This confirms that higher model complexity alone does not guarantee improved accuracy. It is noteworthy that the LSTM performs best on items where the ARIMA differencing parameter d equals zero or one, indicating stable or only mildly nonstationary behavior, suggesting that learning-based models benefit most when temporal structure is largely preserved.

Overall, the table demonstrates that no single forecasting method is optimal for all item types. The results highlight a clear dependence of model performance on data characteristics, with learning-based approaches excelling on highly structured series and classical statistical methods proving more reliable in noisy or weakly structured scenarios.

6. Discussion

The results of this study demonstrate that forecasting effectiveness is strongly determined by the structural properties of the underlying time series. The applied LSTM framework performs exceptionally well when clear periodicity, low noise, and stable temporal patterns are present, confirming that learning-based models can generalize repeating dynamics even from relatively limited historical data. In contrast, for weakly structured or highly volatile series dominated by outliers, LSTM struggles to form a stable representation, and simpler statistical methods, particularly Holt exponential smoothing, remain more robust and reliable. This pattern is well aligned with findings from large-scale forecasting benchmarks and comparative studies, which consistently report that the relative advantage of learning-based models depends strongly on the presence of stable and repeatable temporal structure. Evidence from forecasting competitions and industrial case studies indicates that neural network models tend to outperform classical approaches primarily when seasonality, trend persistence, or deterministic repetition is clearly embedded in the data, while simpler exponential smoothing techniques often remain competitive or superior for noisy and weakly structured series. The present results therefore reinforce the view that forecasting performance is driven more by data characteristics than by model class alone.

Several limitations should be acknowledged. The dataset is limited in size and scope, and a single unified LSTM configuration was applied without extensive hyperparameter optimization or deeper architectures. In addition, the forecasting horizon is relatively long for unstructured series, which amplifies uncertainty and degrades accuracy. External explanatory variables were also not incorporated, although they may significantly influence real-world behavior. From a learning-theoretical perspective, the observed behavior of the LSTM model on highly volatile series can be interpreted as a regression-to-the-mean effect. When trained using a squared-error loss function on data with a low signal-to-noise ratio, neural networks tend to suppress extreme fluctuations and converge toward smoothed representations of the conditional mean. This effect is further amplified in small-sample settings, where limited historical depth restricts the model's ability to distinguish between persistent structure and stochastic variation. As a consequence, the LSTM forecasts systematically underestimate peaks and overestimate troughs, leading to poor explanatory power despite stable numerical convergence.

These findings also have direct implications for operational decision support. In planning environments such as procurement, production scheduling, and inventory control, forecast robustness and stability can be more valuable than marginal improvements in point accuracy. While learning-based models may offer superior performance for highly structured items, their sensitivity to noise can introduce additional uncertainty in downstream decisions. In contrast, classical methods such as Holt exponential smoothing, despite their simplicity, often produce smoother and more predictable forecasts that can lead to more stable inventory and capacity decisions when demand patterns are irregular. Based on these findings, several practical improvements are suggested. Forecast performance for noisy series could benefit from stronger preprocessing, including noise filtering, outlier handling, or smoothed representations. More expressive architectures, such as CNN–LSTM hybrids, attention-based models, or Transformer variants tailored for time series, may further enhance learning. Shorter forecasting horizons would likely improve accuracy for

irregular data, and item-specific model configurations could better capture heterogeneous demand patterns.

The comparative analysis suggests that no single forecasting approach can be considered universally optimal. Instead, effective forecasting requires aligning model choice and configuration with observable data properties such as regularity, volatility, and noise intensity. Learning-based models provide substantial benefits when these properties support pattern generalization, while classical statistical methods remain indispensable baselines for poorly structured and highly uncertain series. Overall, the study demonstrates that LSTM models are highly effective for time series with predictable structure, but their performance is inherently limited when noise and irregular fluctuations dominate. These results highlight the importance of aligning model choice and configuration with measurable data characteristics rather than relying on model complexity alone.

7. Conclusions

The quantitative evaluation confirms that forecasting methods differ substantially in how accurately they reproduce observed weekly values, and these differences can be interpreted consistently using relative deviations. For time series with strong regularity, clear periodic behavior, and low noise, the LSTM model achieves near zero mean squared error and coefficients of determination close to one. In practical terms, this corresponds to an average weekly deviation of only a few percent, even when results are aggregated over the full ten-week evaluation horizon. These results confirm that forecasting accuracy gains achieved by learning-based models are not uniform across all demand patterns. Instead, they emerge primarily when the underlying time series exhibits sufficient regularity, persistence, or periodic repetition to support generalization. In such cases, the LSTM model is capable of learning stable temporal dependencies and delivering highly accurate forecasts even under limited data availability. This finding supports the growing consensus in the forecasting literature that model effectiveness is conditional rather than universal.

For classical statistical methods, deviations are generally larger. Simple baselines such as naïve and moving average forecasting typically result in average weekly deviations in the range of 10 to 20 percent, which accumulate proportionally across the evaluated weeks. Simple exponential smoothing reduces these deviations moderately, while Holt exponential smoothing achieves the lowest errors among the statistical approaches, often reaching single-digit percentage deviations per week for moderately structured series. The consistently strong performance of Holt exponential smoothing across moderately structured and noisy series highlights the continued relevance of classical statistical forecasting methods in operational contexts. Despite their simplicity, these approaches provide stable and interpretable forecasts that are less sensitive to extreme observations and random fluctuations. This robustness makes them particularly suitable for short-term planning tasks, where forecast stability can be as important as point accuracy.

For highly irregular and noise-dominated series, all methods exhibit substantial discrepancies. In these cases, Holt exponential smoothing consistently limits the deviation more effectively than ARIMA or the LSTM model. ARIMA, despite its higher complexity, does not provide a systematic accuracy advantage over simpler statistical techniques. The results further indicate that increased model complexity does not automatically translate into improved forecasting performance. Although ARIMA models incorporate a richer linear structure than exponential smoothing, their accuracy advantages remain limited in highly volatile settings. This observation reinforces the principle that forecasting models should be selected based on data properties and decision requirements rather than theoretical sophistication alone.

From a practical perspective, the findings suggest several actionable guidelines. Learning-based models such as LSTM are most appropriate for items with strong and repeatable temporal structure, where their ability to capture nonlinear dependencies provides a clear advantage. For items dominated by noise or irregular fluctuations, simpler statistical methods remain preferable due to their robustness and transparency. Consequently, forecasting systems should incorporate data-driven criteria for model selection rather than relying on a single method across all items. Overall, the results show that learning-based models deliver substantially lower total error when a strong and stable temporal structure is present. Within the examined dataset, the LSTM model performs excellently on two items and very well on two additional items out of five, demonstrating that it can achieve low percentage deviations in most cases when sufficient regularity exists. At the same time, the findings reinforce that classical statistical methods remain more suitable for series dominated by noise or weak structure.

Author Contributions: Conceptualization, Z.B. and P.V.; methodology, Z.B.; software, Z.B. and P.V.; validation, Z.B., P.V. and Á.B.; formal analysis, Z.B.; investigation, Z.B. and P.V.; resources Z.B. and P.V.; data curation, Z.B. and P.V.; writing—original draft preparation, Z.B., P.V. and Á.B.; writing—review and editing, Z.B., P.V. and Á.B.; visualization, Z.B. and P.V.; supervision, P.V.; project administration, Á.B. All authors have read and agreed to the published version of the manuscript.

Funding: The creation of this scientific communication was supported by the University of Miskolc with funding granted to the author Peter Veres within the framework of the institution’s Scientific Excellence Support Program. (Project identifier: ME-TKTP-2025-067).

Data Availability Statement: The datasets and programming files presented in this study are available at https://github.com/Strigula93/LSTM_PO_FC, accessed on 1 January 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kaur, J.; Parmar, K.S.; Singh, S. Autoregressive models in environmental forecasting time series: A theoretical and application review. *Environ. Sci. Pollut. Res.* **2023**, *30*, 19617–19641. [CrossRef]
2. Khashei, M.; Bijari, M.; Raissi Ardali, G.A. Hybridization of autoregressive integrated moving average (ARIMA) with probabilistic neural networks (PNNs). *Comput. Ind. Eng.* **2012**, *63*, 37–45. [CrossRef]
3. Lee, H.L.; Padmanabhan, V.; Whang, S. Information Distortion in a Supply Chain: The Bullwhip Effect. *Manag. Sci.* **1997**, *43*, 546–558. [CrossRef]
4. Prak, D.; Teunter, R.; Syntetos, A. On the calculation of safety stocks when demand is forecasted. *Eur. J. Oper. Res.* **2017**, *256*, 454–461. [CrossRef]
5. Vandeput, N. *Demand Forecasting Best Practices*; Simon and Schuster: New York, NY, USA, 2023.
6. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. *Int. J. Forecast.* **2020**, *36*, 54–74. [CrossRef]
7. Hewamalage, H.; Ackermann, K.; Bergmeir, C. Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Min. Knowl. Discov.* **2023**, *37*, 788–832. [CrossRef] [PubMed]
8. Fildes, R.; Goodwin, P.; Lawrence, M.; Nikolopoulos, K. Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *Int. J. Forecast.* **2009**, *25*, 3–23. [CrossRef]
9. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 3rd ed.; OTexts: Melbourne, Australia, 2021.
10. Momin, B.; Chavan, G. Univariate Time Series Models for Forecasting Stationary and Non-stationary Data: A Brief Review. In *Smart Innovation, Systems and Technologies*; Springer International Publishing: Cham, Switzerland, 2017; pp. 219–226. [CrossRef]
11. Abdullah, M.M. Using the Single-Exponential-Smoothing Time Series Model under the Additive Holt-Winters Algorithm with Decomposition and Residual Analysis to Forecast the Reinsurance-Revenues Dataset. *Pak. J. Stat. Oper. Res.* **2024**, *20*, 311–340. [CrossRef]
12. Cheng, S.; Quilodrán-Casas, C.; Ouala, S.; Farchi, A.; Liu, C.; Tandeo, P.; Fablet, R.; Lucor, D.; Iooss, B.; Brajard, J.; et al. Machine Learning With Data Assimilation and Uncertainty Quantification for Dynamical Systems: A Review. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1361–1387. [CrossRef]

13. Kristian, A.; Sumarsan Goh, T.; Ramadan, A.; Erica, A.; Visiana Sihotang, S. Application of AI in Optimizing Energy and Resource Management: Effectiveness of Deep Learning Models. *Int. Trans. Artif. Intell.* **2024**, *2*, 99–105. [[CrossRef](#)]
14. Chudo, S.B.; Terdik, G. Modeling and Forecasting Time-Series Data with Multiple Seasonal Periods Using Periodograms. *Econometrics* **2025**, *13*, 14. [[CrossRef](#)]
15. Khashei, M.; Mahdavi Sharif, B. A Kalman filter-based hybridization model of statistical and intelligent approaches for exchange rate forecasting. *J. Model. Manag.* **2020**, *16*, 579–601. [[CrossRef](#)]
16. Alanazi, B.S. A Comparative Study of Traditional Statistical Methods and Machine Learning Techniques for Improved Predictive Models. *Int. J. Anal. Appl.* **2025**, *23*, 18. [[CrossRef](#)]
17. Maragkos, N.; Refanidis, I. A Comparative Evaluation of Time-Series Forecasting Models for Energy Datasets. *Computers* **2025**, *14*, 246. [[CrossRef](#)]
18. Lecaros, A.; Moquillaza, A.; Falconi, F.; Aguirre, J.; Tapia, A.; Paz, F. Selection and Modeling of a Formal Heuristic Evaluation Process Through Comparative Analysis. In *Lecture Notes in Computer Science*; Springer International Publishing: Cham, Switzerland, 2022; pp. 28–46. [[CrossRef](#)]
19. Rangineni, S.; Bhanushali, A.; Suryadevara, M.; Venkata, S.; Peddireddy, K. A Review on Enhancing Data Quality for Optimal Data Analytics Performance. *Int. J. Comput. Sci. Eng.* **2023**, *11*, 51–58. [[CrossRef](#)]
20. Akhtar, S.; Shahzad, S.; Zaheer, A.; Ullah, H.S.; Kilic, H.; Gono, R.; Jasiński, M.; Leonowicz, Z. Short-Term Load Forecasting Models: A Review of Challenges, Progress, and the Road Ahead. *Energies* **2023**, *16*, 4060. [[CrossRef](#)]
21. Ubal, C.; Di-Giorgi, G.; Contreras-Reyes, J.E.; Salas, R. Predicting the Long-Term Dependencies in Time Series Using Recurrent Artificial Neural Networks. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1340–1358. [[CrossRef](#)]
22. Mohammed, N.A.; Al-Bazi, A. An adaptive backpropagation algorithm for long-term electricity load forecasting. *Neural Comput. Appl.* **2021**, *34*, 477–491. [[CrossRef](#)] [[PubMed](#)]
23. Yunita, A.; Pratama, M.I.; Almuzakki, M.Z.; Ramadhan, H.; Akhir, E.A.P.; Firdausiah Mansur, A.B.; Basori, A.H. Performance analysis of neural network architectures for time series forecasting: A comparative study of RNN, LSTM, GRU, and hybrid models. *MethodsX* **2025**, *15*, 103462. [[CrossRef](#)]
24. Kabir, M.R.; Bhadra, D.; Ridoy, M.; Milanova, M. LSTM–Transformer-Based Robust Hybrid Deep Learning Model for Financial Time Series Forecasting. *Sci* **2025**, *7*, 7. [[CrossRef](#)]
25. Hasan, M.R.; Islam, M.R.; Rahman, M.A. Developing and implementing AI-driven models for demand forecasting in US supply chains: A comprehensive approach to enhancing predictive accuracy. *Edelweiss Appl. Sci. Technol.* **2025**, *9*, 1045–1068. [[CrossRef](#)]
26. Li, X.; Xiong, H.; Li, X.; Wu, X.; Zhang, X.; Liu, J.; Bian, J.; Dou, D. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowl. Inf. Syst.* **2022**, *64*, 3197–3234. [[CrossRef](#)]
27. Khudhair, Z.S.; Zubaidi, S.L.; Ortega-Martorell, S.; Al-Ansari, N.; Ethaib, S.; Hashim, K. A Review of Hybrid Soft Computing and Data Pre-Processing Techniques to Forecast Freshwater Quality's Parameters: Current Trends and Future Directions. *Environments* **2022**, *9*, 85. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.