Subject: Response Letter – Submission ID 249763825

Dear Professor Dolgui,

Please find attached the revised version of our manuscript submitted to the *International Journal of Production Research*.

We sincerely appreciate the insightful and constructive comments provided by the reviewers. We have carefully considered each suggestion and made substantial revisions to improve the clarity, depth, and overall quality of the manuscript.

Below is a summary of the key revisions made in response to the reviewers' feedback:

- Provided a clearer and more detailed explanation of the probabilistic forecasting approach.
- Corrected mathematical notations throughout the manuscript.
- Clarified concerns raised regarding the dataset used in the study.
- Edited and refined the MSSE and CRPS equations.
- Added a table of summary statistics.
- Added tables presenting computational time , in addition to accuracy.
- Clarify the setup of the experiment for LSTM, and updated the LSTM result and discussion.
- Expanded the discussion section to elaborate on how the forecasts can be applied in practice and their implications for pharmaceutical supply planning in Ethiopia.
- Organized and updated the list of references.
- Clearly explained how domain knowledge was integrated into the forecasting models, including the selection and validation of expert-informed predictors.
- Included actionable recommendations for applying the study findings in real-world contexts and included a new subsection on managerial implications.
- Updated future research direction.

We believe these revisions have significantly strengthened the manuscript, and we hope it now meets the expectations of the journal and the reviewers. All changes have been highlighted in blue for your convenience in both the response letter and the revised manuscript.

Thank you for your time and consideration.

Sincerely,

Reviewer 1

Summary

In the paper, the authors study the improvement on the forecasting performance by using domain-specific knowledge for Ethiopian Pharmaceutical Supply Service (EPSS). They found that integrating expert-identified variables (stock replenishment schedules, fiscal inventory counts, and disease outbreaks) can largely improve the forecasting performance. Additionally, they explore probabilistic forecasting to some extent.

General Comments

Comment 1: Overall, I am modestly positive about the work, given the importance of pharmaceutical forecasting in low-income countries. However, I would like the authors to address a number of concerns and issues which I think it will improve its quality while also enhancing the understanding of the readers of IJPR.

Response: Thank you for taking the time to review our paper. We appreciate your positive comments and critical evaluation. We have addressed your comments in the following responses.

Specific Comments

Comment 1: In the current version, the probabilistic forecasting part should be more elaborated. There is only a brief sentence on how their probabilistic forecasting is generated (i.e. bootstrapping). Are there alternative ways to generate probabilistic forecasting? Why did authors choose to bootstrap?

Response: Thank you for highlighting the need for a more detailed explanation of our probabilistic forecasting approach. We agree that a deeper discussion on this point would strengthen the clarity of our study. In response, we have expanded the corresponding section of the manuscript by adding the following section.

In addition to point forecasts, we generated probabilistic forecasts to capture the uncertainty surrounding future pharmaceutical consumption. Several approaches are available for generating probabilistic forecasts, including analytical prediction intervals, quantile regression, Bayesian modeling through Markov Chain Monte Carlo (MCMC) methods, bootstrapping, and conformal prediction (Wang et al., 2023).

In this study, we employed a bootstrapping approach to construct predictive intervals. Bootstrapping was chosen primarily for its flexibility and model-agnostic nature, allowing it to be applied uniformly across the diverse range of forecasting methods implemented without requiring strong distributional assumptions. Moreover, pharmaceutical consumption data often exhibit irregular and volatile patterns, making non-parametric approaches like bootstrapping particularly suitable.

Specifically, we assume that future forecast errors will be similar to past forecast errors. The forecast error at time t is defined as: e_t = y_t - $\hat{y}$_t where y_t represents the observed consumption, and $\hat{y}$_t denotes the corresponding forecast estimate. To simulate future consumption paths, we randomly sample errors with replacement from the historical error

distribution and add them to the point forecast estimates. This process is repeated multiple times (1,000 iterations in our study) to generate a distribution of possible outcomes for each forecast horizon.
Prediction intervals at the desired confidence level (e.g., 95%) are then constructed by taking appropriate quantiles from the empirical distribution of the simulated forecasts (Hyndman & Athanopolous, 2021)

The bootstrapping method thus provides a robust and flexible framework to quantify forecast uncertainty across a heterogeneous set of pharmaceutical products without imposing restrictive parametric assumptions.

Comment 2: At the moment, Section 4.1 only provides a brief description of probabilistic forecasting. More importantly, how to take probabilistic forecasting into account for the decision- making is not elaborated. I suggest that the authors have a simple numerical example (e.g. Wang et al. 2023) to showcase the value of probabilistic forecasting.

Response: Thank you for this insightful suggestion. We agree that providing a concrete example of how probabilistic forecasts can be applied in decision-making would help clarify their practical value. We note that the main focus of this paper is to evaluate the integration of domain knowledge—gathered through interviews with experts from practical supply service settings—into the modeling process. A thorough analysis of how forecasts, including both point forecasts and probabilistic forecasts, can be used would require a detailed understanding of inventory policies and the development of simulation models to assess how such forecasts inform decision-making. However, this lies beyond the scope of the specific question we aim to address in this paper. We believe future research could explore this direction in greater depth. However, in response, we have expanded Section 4.1 of the manuscript to include a simple illustrative example that shows how a manager might use a prediction interval to inform inventory decisions. The following section has been added in the revised manuscript:

To further illustrate the practical relevance, we consider a case where the mean forecasted consumption for the next month is 1,000 units, with a 90% prediction interval ranging from 850 to 1,150 units. Table 1 summarizes the decision outcomes under each approach.

Table 1: Comparison of ordering decisions under point and probabilistic forecasts - an illustrative example.

Forecast Type	Ordering Quantity	Risk Consideration
Point Forecast Only	1,000 units	No explicit consideration of uncertainty
Probabilistic Forecast (95% service level)	1,150 units	Adjusts inventory to account for demand variability

Under a traditional approach, inventory decisions would rely solely on the point forecast. Based on the mean prediction of 1,000 units, the store manager would order precisely that quantity, assuming it would meet the expected demand. However, this approach does not incorporate any adjustment for uncertainty, potentially leading to stockouts if actual demand exceeds 1,000 units, or excess inventory if consumption is lower.

In contrast, utilizing the full probabilistic forecast allows the decision-maker to take

variability into account. If a higher service level is required, for example 95%, the inventory policy may be adjusted by stocking closer to the upper bound of the 90% prediction interval (e.g., 1,150 units). If inventory holding costs are a major concern and the organization can tolerate a modest risk of shortage, the order quantity could remain closer to the median forecast of 1,000 units.

This illustrative comparison indicates that point forecasts provide a single estimate without adjusting for forecast uncertainty, while probabilistic forecasts allow decision-makers to explicitly align inventory decisions with risk tolerance and service level requirements. This is particularly important for pharmaceutical products, where the consequences of stockouts or overstocking can be significant both operationally and clinically.

Comment 3: Additionally, I also encourage authors to demonstrate how to use point forecasting for the decision-making as well. And then comparing the difference in the decision-making by point forecasting and probabilistic forecasting. In this way, the value of probabilistic forecasting can be clearly highlighted.

Response: Thank you for your valuable suggestions. We agree that demonstrating how point forecasts and probabilistic forecasts can be used for decision-making, along with a comparison, would strengthen the practical contribution of the paper. We also provide a response to comment 2, which describes how forecasts might be used, and as also discussed in the previous comment, we believe that further future investigation is required to better understand how forecasts, especially probabilistic forecasts, can be integrated into inventory policy.

We have added the following text to the conclusion section of the revised manuscript:

Future research should explore the integration of probabilistic forecasting into inventory management policies, with a focus on evaluating the operational impact of forecast uncertainty. In addition, there is substantial potential in leveraging more granular data—such as daily or weekly consumption patterns—and examining forecasts at hierarchical levels, including individual sites or health facilities. Investigating how fine-grained temporal and spatial forecasts influence decision-making in inventory control could provide valuable insights into reducing stockouts, minimizing holding costs, and decreasing pharmaceutical waste.

We also included a new section where we discuss the running time of each model:

Computational efficiency and resource considerations

In addition to forecast accuracy, it is also important to consider the computational efficiency of each model, particularly in settings where computational resources and technical expertise are limited.

Table 3 shows the total runtime required to train and generate forecasts across all 33 pharmaceutical product time series for each method for all origins. All models were implemented using R, except TimeGPT, which was run via Google Colab using a T4 GPU backend. All other models were executed on a local machine with 7 CPU cores (11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40 GHz and 8 GB RAM).

Table 3: Computation time required for training and generating forecasts for each model across all products.

Model	Runtime (Seconds)	Runtime type	category
LSTM	6892.30	CPU with 7 cores	high
LSTM with regressors	7234.00	CPU with 7 cores	high
Regression	27.48	CPU with 7 cores	low
Regression with regressors	47.38	CPU with 7 cores	low
ARIMA	96.27	CPU with 7 cores	low
ARIMA with regressors	138.08	CPU with 7 cores	medium
TimeGPT	2.57	Colab T4 GPU	low
TimeGPT with regressors	3.72	Colab T4 GPU	low
ETS	75.71	CPU with 7 cores	low

As shown in Figure 4 and 5, models that incorporated expert-informed regressors generally achieved better forecast accuracy compared to their univariate versions. This trend was most evident in classical models such as ARIMA and regression, where the inclusion of regressors led to a noticeable shift toward lower and more concentrated error distributions. However, this improvement came with an increase in computational cost. In all cases, the addition of regressors increased runtime—by 72% in the regression model (from 27.48 to 47.38 seconds) and by 43% in ARIMA (from 96.27 to 138.08 seconds). TimeGPT also showed a slight increase in runtime (from 2.57 to 3.72 seconds), although the total processing time remained extremely low overall. Moreover, LSTM exhibited the largest increase in runtime when regressors were added, rising from 6,892 to 7,234 seconds. This sharp increase reflects the sensitivity of neural networks to input configuration, especially in contexts with limited data, irregular demand, and noisy signals.

These results underscore the importance of balancing model performance with implementation cost. While deep learning models such as LSTM offer strong potential when well-tuned, they demand significantly more computational resources and may be less robust when integrating static or weakly aligned contextual features. By contrast, TimeGPT, a foundational model pretrained on large-scale time series data, provides a compelling alternative. It required less than 4 seconds to forecast all products, offered competitive accuracy, and required no tuning or retraining—making it well-suited for practical use in low-resource environments.

Moreover, we have expanded Section 4.1 to illustrate how decision-makers would traditionally use a point forecast for inventory decisions and how incorporating probabilistic forecasts enables more risk-informed planning using an illustrative example.

To further illustrate the practical relevance, we consider a case where the mean forecasted consumption for the next month is 1,000 units, with a 90% prediction interval ranging from 850 to 1,150 units.

Under a traditional approach, inventory decisions would rely solely on the point forecast. Based on the mean prediction of 1,000 units, the store manager would order precisely that quantity, assuming it would meet the expected demand. However, this approach does not incorporate any adjustment for uncertainty, potentially leading to stockouts if actual demand exceeds 1,000 units, or excess inventory if consumption is lower.

In contrast, utilizing the full probabilistic forecast allows the decision-maker to take variability into account. If a higher service level is required, for example 95%, the inventory policy may be adjusted by stocking closer to the upper bound of the 90% prediction interval (e.g., 1,150 units). If inventory holding costs are a major concern and the organization can tolerate a modest risk of shortage, the order quantity could remain closer to the median forecast of 1,000 units.

Table 1: Comparison of ordering decisions under point and probabilistic forecasts - an illustrative example.

Forecast Type	Ordering Quantity	Risk Consideration
Point Forecast Only	1,000 units	No explicit consideration of uncertainty
Probabilistic Forecast (95% service level)	1,150 units	Adjusts inventory to account for demand variability

This illustrative comparison indicates that point forecasts provide a single estimate without adjusting for forecast uncertainty, while probabilistic forecasts allow decision-makers to explicitly align inventory decisions with risk tolerance and service level requirements. This is particularly important for pharmaceutical products, where the consequences of stockouts or overstocking can be significant both operationally and clinically.

Comment 4: The mathematical notations on Page 9-10 are not properly shown.
Response: Thank you, this is now corrected.

Reviewer 2

Comment 1: The study looks at a pharmaceutical supply chain and develops point and probabilistic forecasting models for a case study in Ethiopia. A number of models have been developed and tested rigorously. Please see below several comments that authors may find useful.

Response: Thank you for taking the time to review our paper. We appreciate your positive comments and critical evaluation. We have addressed your comments in the following responses.

Comment 2: I think the dataset is small, with only 33. I understand the limitations when it comes to real cases. Is it possible to replicate this with bigger data? This is an empirical study and you can use open source data such as Rossman store sales.

Response 2: Thank you, we understand the concern about the dataset, but please let us clarify the context and why we think this is relevant.

The dataset used in our study was selected based on the recommendations of the Ministry of Health of Ethiopia, which identifies 25 essential pharmaceuticals frequently used in national-

level assessments. These medicines are widely recognized as critical to the country's healthcare system and are considered representative of broader pharmaceutical demand. To further enhance the relevance and breadth of our analysis, we included an additional eight pharmaceuticals beyond the official list, bringing the total to 33. We believe this sample size is appropriate for demonstrating the study's primary objective: understanding how domain knowledge can be effectively integrated into forecasting models. While publicly available datasets—such as the Rossmann store sales data—can be useful for exploring general forecasting techniques, they are not suitable for the aims of this study. The Rossmann dataset, for example, reflects retail sales patterns in a commercial setting and does not capture the complexities of pharmaceutical supply chains. These include regulatory constraints, essential medicine prioritization, demand volatility due to epidemiological factors, and health system-specific procurement and distribution dynamics. Our goal in this paper is not simply to evaluate predictive performance but to investigate how expert domain knowledge can be systematically incorporated into modeling processes relevant to public health supply chains. To that end, our collaboration with the Ethiopian Pharmaceuticals Supply Service (EPSS) provided access to subject matter experts who contributed deep contextual insights. These experts informed key aspects of the study, including the identification of relevant medicines, domain knowledge and insight on what affects the consumption of products in the country. This kind of integration between qualitative domain knowledge and quantitative modeling would not be feasible using generic, publicly available datasets.

Comment 2: Readers will appreciate some exploratory analysis on data. Figure 2 shows some products, but are they representative? Perhaps you could explore the features of data, and some summary statistics, given that it is not possible to plot all of them here.

Response: Thank you for the comment. We have provided two main features of data to highlight the strength of trends and seasonality in Figure 1. Additionally, and to address your comment about adding summary statistics, we have now included a table to provide several summary statistics.

Table also provides the summary statistics computed for each pharmaceutical product time series include traditional distributional measures—mean, median, standard deviation, minimum, and maximum—alongside time series–specific and structural characteristics. These include the first-order autocorrelation (ACF at lag 1), the mean length of consecutive zero runs (zero_run_mean), and the squared coefficient of variation computed only on non-zero values (nonzero_squared_cv). Together, these metrics capture not only central tendency and dispersion, but also temporal dependence, sparsity patterns, and relative variability in periods with non-zero consumption.

Table 2: Summary statistics of the pharmaceutical product time series

item	Mean	Median	Standard deviation	Miinum	Maximum	ACF lag 1	zero_run_mean	nonzero_squared_cv
Adrenaline (Epinephrine) - 0.1% in 1ml Ampoule - Injection	83077.42	69800.00	55257.59	0	204800	0.68	1.0	0.42
Amlodipine - 5mg - Tablet	330051.67	84350.00	473283.80	0	1495700	0.82	2.0	1.75
Amoxicillin - 500mg - Capsule	25880.97	25291.00	15751.99	1137	75305	0.30	0.0	0.37
Anti-Rho (D)	1911.10	1983.00	1475.71	30	8477	0.11	0.0	0.60
Artemether + Lumefanthrine	7364.28	953.00	17714.45	0	88469	0.64	6.5	4.33
Atenolol - 50mg - Tablet	3695.75	3312.50	2587.84	4	12661	0.27	0.0	0.49
Atrovastatin - 20mg - Tablet	3618.97	1302.50	5506.25	0	27973	0.44	1.0	2.26
Ceftriaxone	542999.03	553782.50	346067.19	0	1285708	0.47	2.0	0.36
Dextrose	312489.42	296130.00	255750.02	0	1004850	0.53	5.0	0.39
Frusemide - 10mg/ml in 2ml Ampoule - Injection	176436.17	175360.00	137620.74	17880	940930	0.34	0.0	0.61
Gentamicin	201360.50	182040.00	181410.03	0	654000	0.64	2.0	0.75
Hydralazine - 20mg/ml in 1ml Ampoule - Injection	2952.17	2456.00	2741.23	6	15713	0.57	0.0	0.86
Insulin Isophane Human(Suspension)	66717.83	73816.00	46364.04	0	155631	0.60	11.0	0.21
Insulin Soluble Human	8973.33	8890.50	7140.69	0	28443	0.53	11.0	0.33
Insuline Isophane Biphasic	12795.87	11402.50	8475.27	860	38586	0.45	0.0	0.44
Lamivudine + Efavirenz +Tenofovir	145178.95	167890.50	133576.98	0	550780	0.61	11.0	0.51
Lamivudine + Zidovudine	26784.80	22579.00	22565.25	0	89910	0.55	11.0	0.39
Lidocaine HCL	80100.70	62702.00	105600.48	1000	791050	0.12	0.0	1.74
Magnesium Sulphate	3573.83	1832.50	4465.78	0	22649	0.49	11.0	1.09
Medroxyprogesterone	501663.00	560692.00	328276.14	0	984774	0.65	11.0	0.16
Metformin - 500mg - Tablet	26505.25	20294.75	21695.60	654	86705	0.42	0.0	0.67
Omeperazole - 20 mg - Capsule (Enclosing Enteric Coated Granules)	46368.55	40012.00	38379.32	0	176219	0.62	6.0	0.52
Omeperazole - 4mg/ml in 10ml - Injection	11438.88	3345.50	17643.52	0	81355	0.55	5.0	1.82
Oral Rehydration Salt	507220.42	558426.00	394581.24	0	1384257	0.70	6.0	0.28
Oxytocin	180155.23	151480.00	164309.63	0	846060	0.32	11.0	0.49
Pentavalent	422071.73	474297.50	238940.77	0	944819	0.78	11.0	0.08
Propylthiouracil - 100mg - Tablet	7044.46	6426.50	6112.89	0	28473	0.44	2.0	0.69
RHZ (Rifampicin + Isoniazid + Pyrazinamide)	2260.05	1759.00	2233.00	0	7095	0.70	8.0	0.45
Rapid Diagnostic Test	377843.65	360187.50	264551.71	0	1004325	0.63	11.0	0.21
Ringer's Injection	88896.27	95227.00	54255.34	390	251283	0.53	0.0	0.37
Sodium Chloride (Normal Saline)	300706.82	289171.50	136772.04	60759	702526	0.36	0.0	0.21
Sulphamethoxazole + Trimethoprim - (200mg + 40mg)/5ml - Suspension	84216.90	65917.50	72963.22	20	259512	0.59	0.0	0.75
Tetracycline - 1% - Eye Ointment	199370.65	172721.50	139041.86	7556	585987	0.02	0.0	0.49

Comment 3: Page 4, line 13, a reference is missing as shown in question mark.

Response: Thank you for identifying the error, and we have corrected the mistakes accordingly.

Comment 4: In page 9, before section 3.3 you describe how you generated probabilistic forecasts. Since you have claimed probabilistic forecasting is one of your innovations, I would expect a more detailed description for readers.

Response: Thank you for highlighting the need for a more detailed explanation of our probabilistic forecasting approach. We agree that a deeper discussion on this point would strengthen the clarity of our study. In response, we have expanded the corresponding section of the manuscript by adding the following section.

In addition to point forecasts, we generated probabilistic forecasts to capture the uncertainty surrounding future pharmaceutical consumption. Several approaches are available for generating probabilistic forecasts, including analytical prediction intervals, quantile regression, Bayesian modeling through Markov Chain Monte Carlo (MCMC) methods, bootstrapping, and conformal prediction (Wang et al., 2023).
In this study, we employed a bootstrapping approach to construct predictive intervals. Bootstrapping was chosen primarily for its flexibility and model-agnostic nature, allowing it to be applied uniformly across the diverse range of forecasting methods implemented without requiring strong distributional assumptions. Moreover, pharmaceutical consumption data often exhibit irregular and volatile patterns, making non-parametric approaches like bootstrapping particularly suitable.
Specifically, we assume that future forecast errors will be similar to past forecast errors. The forecast error at time t is defined as: e_t = y_t - $\hat{y}$_t where y_t represents the observed consumption, and $\hat{y}$_t denotes the corresponding forecast estimate. To simulate future consumption paths, we randomly sample errors with replacement from the historical error distribution and add them to the point forecast estimates. This process is repeated multiple times (1,000 iterations in our study) to generate a distribution of possible outcomes for each forecast horizon.

Prediction intervals at the desired confidence level (e.g., 95%) are then constructed by taking appropriate quantiles from the empirical distribution of the simulated forecasts (Hyndman & Athanopolous, 2021)

The bootstrapping method thus provides a robust and flexible framework to quantify forecast uncertainty across a heterogeneous set of pharmaceutical products without imposing restrictive parametric assumptions.

Comment 5: MSSE and CRPS equations on page 10 needs rewriting.

Response: this is now corrected.

Comment 6: As we can see in Figure 3, lstm has generated some bad forecasts. This is evident in MASE, some being over 40. What is the reason? What are these products (I appreciate this is explained on page 12, second paragraph, but there are several of them in this data, and you may exclude them from the analysis to have a fair comparison. Having said that, I think you can manage these forecasts with better training lstm and rigorous regularization. This graph is not informative because it is impacted by outliers, and we can't see the distribution for the other items.

Response: Thank you for your feedback regarding the LSTM forecasts.

Following your suggestions, we undertook several improvements to enhance the robustness and fairness of the LSTM modeling. Firstly, we identified that certain products exhibited extremely high MASE values, largely due to the challenges of forecasting small-volume or highly volatile demand patterns.

To address model overfitting, we revised our LSTM setup to include dropout regularization and implemented early stopping based on validation loss. Additionally, we improved the test set alignment to ensure correct forecasting horizons during model evaluation. These changes were applied consistently across both univariate and multivariate LSTM frameworks.

Furthermore, we conducted a preliminary hyperparameter tuning exercise, evaluating combinations of key LSTM settings (e.g., number of units, dense layer size, dropout rates, batch sizes) based on validation performance across a subset of products. Based on this analysis, we selected 50 LSTM units, 100 dense units, a 0.2 dropout rate, and a batch size of 32 as a balance between model complexity and generalization.

Furthermore, with the model fine-tuning, the extreme outlier issue was also solved. We have now provided the revised modelling framework and incorporated the updated figures into the manuscript.

We have added the following text to the section 3:

To improve the robustness of the LSTM models and address overfitting issues, we introduced dropout regularization layers after the LSTM units and employed early stopping based on validation loss during model training. Each LSTM model was trained independently for each product series.

Hyperparameter tuning was performed in a preliminary phase using a subset of products. Various configurations of LSTM units (30–100), dense units (50–200), dropout rates (0.1–0.5), and batch sizes (16–64) were evaluated based on validation set performance. The final model structure — 50 LSTM units, 100 dense units, a 0.2 dropout rate, and a batch size of 32 — was selected as a trade-off between forecast accuracy and model stability across different demand patterns.

We have included the following clarification to the Results and Discussion section:

While LSTM models achieved the best overall forecast accuracy across products, their performance exhibited notable variability depending on the characteristics of individual demand patterns. For example, the product "Amlodipine - 5mg - Tablet" demonstrates periods of extreme variability, with spikes in demand followed by periods of very low or zero consumption. Such patterns align well with the strengths of univariate LSTM models, which are adept at capturing long-term dependencies and managing complex temporal fluctuations. In contrast, the demand for "Anti-Rho (D)" is erratic and sparse, with frequent random fluctuations and little structural consistency. This lack of clear temporal patterns can make it challenging for LSTM models to learn generalizable signals, particularly given the limited data length available for training. Although LSTM can manage irregular data to some extent, it performs best when patterns are consistent or cyclic. These variations across product types contributed to the observed distribution of forecast errors across the product portfolio. Moreover, although we expect that multivariate LSTM models benefit from expert-informed predictors, our results show that the univariate LSTM consistently achieved better forecast performance. This outcome may be attributed to the nature of the predictors—binary, static, or weakly aligned with short-term temporal dynamics—which can disrupt rather than enhance learning when added to a neural network sensitive to input configuration. Moreover, with relatively short historical series and limited training samples, the inclusion of additional variables may have led to overfitting or reduced generalization. These findings suggest that while LSTM models can effectively learn temporal patterns from consumption data alone, incorporating structured external knowledge requires careful feature engineering and alignment to be beneficial.

Comment 7: It would be useful to have tables for accuracy.

Response: Thank you for your comment. We note that the figures displays the distribution of errors using a boxplot, which also conveys central tendency measures. Since these summary statistics are already visually represented, including a separate table would introduce redundancy. Therefore, we have chosen to present these summaries within the figure rather than adding a table.

Comment 8: How are you planning to use these forecasts? What are the implications? This can add to the value of the research.

Response: Thank you for this insightful comment. We agree that clearly articulating the practical use and implications of the forecasts will enhance the value of the research. We have now expanded the discussion section to explain how the forecasts are intended to be used and their implications for pharmaceutical supply planning in Ethiopia.

Managerial Implications

This study offers actionable lessons for supply chain managers, public health planners, and policymakers navigating the uncertainty of pharmaceutical demand—particularly in systems like Ethiopia's, where operational constraints and demand volatility are the norm, not the exception.

At present, national forecasting at the EPSS still relies heavily on basic extrapolation tools—usually Excel sheets or donor-developed software such as the Quantification Analysis Tool (QAT). These tools are easy to use but limited: they assume stable demand, ignore uncertainty, and often miss the operational signals embedded in local experience. In practice, forecasts are sometimes adjusted based on gut feeling or anecdotal program insights—not because planners want to—but because the tools don't offer a better alternative.

The models developed here—freely available, open-source, and designed to integrate directly with routine EPSS data—allow planners to make decisions using full probabilistic forecasts, not just single-point estimates. Forecasts are delivered with prediction intervals (e.g., 80% or 90%) that help teams decide not only how much to order, but how much risk they're willing to tolerate. For products with known seasonality—like antimalarials or diagnostic kits—this matters. The system doesn't just forecast a number; it gives planners a buffer strategy.

Beyond accuracy, the models also reflect how medicines actually move through the system. Predictors based on warehouse replenishments, fiscal year inventory routines, and known disease cycles are embedded into the forecasts—not hard-coded, but learned from the data in ways that are transparent and reproducible. These inputs are drawn from expert operational knowledge and can be updated or modified as the system evolves.

Importantly, the entire modeling pipeline is built in R and Python, and is already shared alongside code and data. This makes it immediately usable by EPSS analysts and regional logistics teams, even in settings without access to proprietary tools or high-end infrastructure. It's not a black box; it's something the system can own, modify, and grow with.

To support real uptake, we recommend five practical steps:

- Embed these probabilistic models into EPSS quantification rounds to replace rigid extrapolation methods and bring scenario-based planning into monthly and quarterly cycles.
- Use prediction intervals to define risk-informed buffer policies, especially for high-volatility products. Not every item needs the same margin of error.
- Automate the collection of domain knowledge (e.g., replenishment events, stock counts, seasonal triggers) into the logistics data stream to reduce reliance on manual elicitation while preserving contextual intelligence.
- Train regional and central teams using the open-source tools and shared codebase, turning forecasting into a hands-on, in-country competency rather than a dependency on external tools or consultants.
- Establish routine forecast review sessions, supported by visual dashboards that reflect not just "what the model says," but how uncertain that estimate is—and what that means for stock levels, procurement plans, and emergency readiness.

Taken together, these steps push forecasting beyond formulas and into real decision-making.

Comment 9: I couldn't find the GitHub repo; is it included in the text?

Response: We have a placeholder in the paper and will include a GitHub repo once the paper is accepted for publication.

Comment 10: The references look unorganized. e.g., line 25, page 17

Response 10: This is now corrected.

Reviewer 3

Comment 1: This manuscript addresses a critical topic in pharmaceutical supply chain management by offering an approach that integrates domain-specific knowledge into forecasting models. While this focus is timely and relevant, the manuscript falls short in several areas, requiring significant revisions to enhance its scientific rigor, methodological clarity, and practical relevance.

Response 1: Thank you for taking the time to review our paper. We appreciate your positive comments and critical evaluation. We have addressed your comments in the following responses.

Comment 2: The study deeply depends on insights from domain experts, such as malaria seasonality, fiscal year inventory periods, and stock replenishment cycles. However, the methodology for incorporating these factors into forecasting models is not well-defined. Furthermore, the approach may not generalize to other pharmaceutical supply chains with differing contextual factors. Thus, the authors should clearly outline the process for translating domain knowledge into actionable predictors within the models. They should also provide a replicable framework or workflow for integrating domain knowledge, ensuring its applicability across different contexts and discuss how this reliance on expert insights can be mitigated or replaced with systematic, scalable data collection methods.

Response 2: Thank you for your thoughtful suggestions, let us briefly discuss the process we followed to identify relevant domain knowledge and its integration into modelling. To systematically integrate expert knowledge into forecasting models in a structured way, we used a multi-step approach involving expert engagement, collaborative data review, data validation, and the transformation of domain insights into model-ready variables. This approach was designed to formalize the incorporation of contextual factors—such as seasonality, inventory cycles, and distribution patterns—into time series forecasting, thereby avoiding reliance on ad hoc assumptions.

The process began in collaboration with the Forecasting and Market Shaping Directorate at Ethiopian Pharmaceutical Supply Service, which helped identify experts with relevant operational expertise. Based on their recommendations, we engaged six professionals from

the Warehouse and Inventory Management Directorate, each with over a decade of experience in pharmaceutical logistics. These experts are actively involved in key processes including inventory movements, order fulfillment, and central-to-regional distribution.

A joint half-day workshop was held with these experts to review five years of monthly consumption data across 33 pharmaceutical products. Through a series of time series visualizations, we collaboratively identified anomalous patterns—such as abrupt spikes, prolonged troughs, and irregular fluctuations. For each anomaly, potential causes were discussed, including planned inventory interventions, emergency distributions, and supply chain disruptions.

To validate these insights, we triangulated the expert input with operational records, including bin cards and warehouse transaction logs. These documents provided detailed information at the batch and transaction level, including timestamps, receipts, and issues. Ethiopian calendar dates were converted to the Gregorian calendar for temporal alignment, and internal stock transfers were excluded to focus on events with external distribution implications.

Based on the validated insights, candidate predictors were proposed by experts to capture observed patterns within a structured modeling framework. A predictor was retained only if all six expert participants agreed it represented a consistent and causally plausible driver of consumption. This consensus-driven procedure helped mitigate individual biases and ensured the predictors reflected shared operational understanding.

The final set of predictors included:

- **Stock Replenishment Events** — Binary indicators representing periods when significant volumes were dispatched from central warehouses.
- **Fiscal Year Inventory Count** — A categorical variable marking the annual physical stock count (typically conducted in July–August), accounting for pre-count stockpiling and temporary distribution halts.
- **Malaria Seasonality** — Dummy variables indicating high-incidence malaria periods (March–May and September–December), capturing known seasonal consumption shifts.

Each predictor was encoded deterministically, with known historical and prospective values, and incorporated into the modeling dataset alongside the raw consumption series. This setup facilitated consistent use across both point forecasts and probabilistic modeling frameworks, while anchoring the modeling process in verifiable operational realities.

We have incorporated the following paragraphs into the manuscript to ensure a clear and thorough description of this process.

To incorporate expert knowledge into time series forecasting in a structured and replicable manner, we developed a multi-step process involving expert engagement, collaborative data review, and operational validation. In partnership with the Forecasting and Market Shaping Directorate at EPSS, we identified ten experienced professionals from the Warehouse and Inventory Management Directorate with deep operational knowledge of pharmaceutical

logistics.

Figure 3 presents a visual overview of the end-to-end approach, highlighting each major stage of the process from expert engagement to model integration.

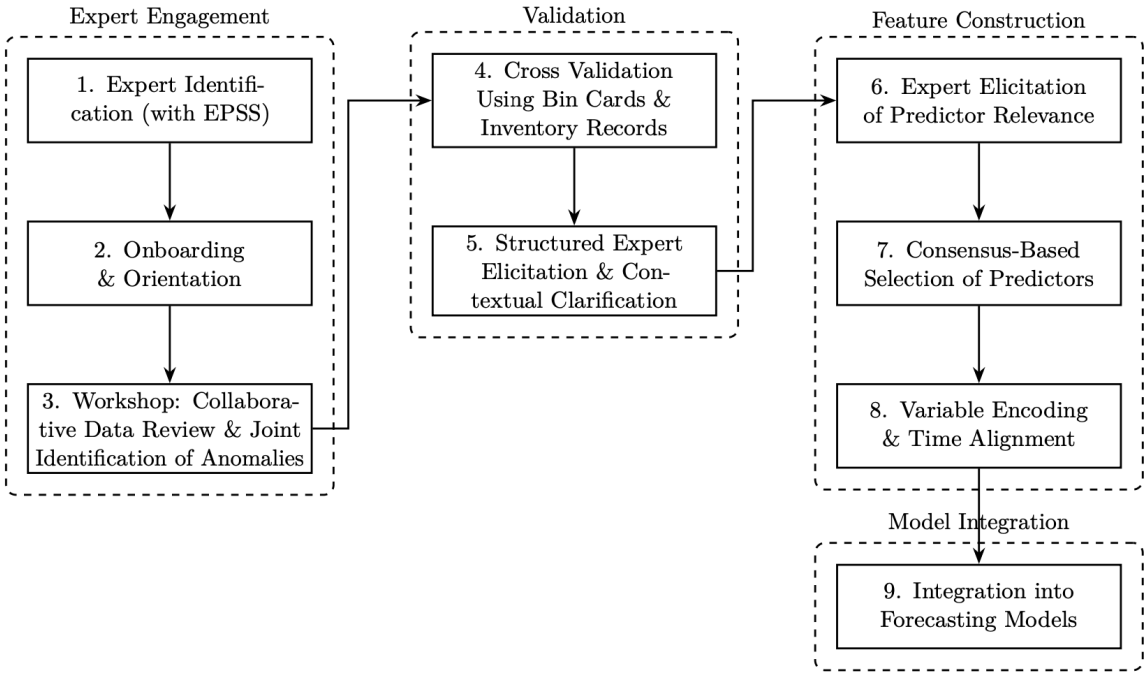

Figure 3: Diagram illustrating the sequential steps involved in collecting, structuring, and utilizing domain knowledge for modeling.

Through a facilitated half-day workshop, these experts reviewed five years of monthly consumption data for 33 pharmaceutical products. Anomalous patterns—such as consumption spikes, prolonged lows, and erratic fluctuations—were jointly identified and interpreted in light of operational events (e.g., emergency distributions, inventory cycles). These insights were cross-validated against bin cards and warehouse records, and internal transfers were excluded to isolate externally relevant events. Dates were standardized to the Gregorian calendar.

Predictor variables were defined through expert consensus, with inclusion contingent on agreement among all ten core participants regarding the operational relevance and causal validity of each factor. The final predictors included binary indicators for stock replenishment events, categorical variables for fiscal year inventory counts, and dummy variables for malaria seasonality. A further description of these variables are summarized as followings:

- Stock replenishment: refers to the process of restocking or refilling inventory to ensure that there are sufficient quantities of products or materials available to meet demand. Whenever there was stock replenishment at the central EPSS, consumption and distribution to hubs and health facilities increased. This increase was attributed to the need to restock depleted inventories and the push from central EPSS to manage space constraints.

- Physical fiscal year inventory counting: refers to the process of manually counting and

verifying the actual quantities of pharmaceutical products available in stock at a specific location. The process is critical for maintaining the accuracy of inventory records, ensuring that medicines are available when needed, and preventing stockouts or overstocking. Physical inventory counting periods also influenced consumption. Stores closed during these periods, halting transactions. We observed increased consumption before inventory counting periods, as hubs and facilities stocked up. July and August were identified as physical counting periods each year.

- Malaria seasonality: Refers to the predictable patterns and fluctuations in malaria incidence throughout the year, typically influenced by climate and environmental conditions. In many regions, malaria transmission peaks during and shortly after the rainy season, when conditions such as stagnant water pools create ideal breeding sites for the Anopheles mosquitoes that transmit the disease. Conversely, malaria cases often decline during the dry season when mosquito breeding sites are reduced. During peak malaria seasons, there is a significant surge in the demand for antimalarial drugs and other related treatments. Malaria seasonality was another significant predictor. Certain pharmaceuticals, like Artemether + Lumefanthrine and Rapid Diagnostic Test kits, were affected by malaria outbreaks. We identified epidemic periods affecting consumption: September to December 2017, March to May 2018, September to December 2018, March to May 2019, September to December 2019, March to May 2020, September to December 2020, March to May 2021, September to December 2021, and March to May 2022.

These predictors, encoded with known historical and future values, were added to the modeling dataset. This approach ensured consistent integration of expert-informed contextual variables across both point and probabilistic forecasting models.

Comment 2: On the other hand, the study's findings are deeply rooted in the Ethiopian Pharmaceutical Supply Service (EPSS) context, with little consideration for generalizability to other healthcare systems or countries. This narrow scope limits the impact and applicability of the research. The authors do not discuss how the proposed methodology can be adapted to other supply chains, particularly those with different logistical challenges. Thus, it is important to identify general principles or frameworks resulting from the study that can be applied to diverse contexts

Response: Thank you. While this study is grounded in the context of the Ethiopian Pharmaceutical Supply Service (EPSS), many of the operational characteristics observed—such as centralized procurement, periodic stock replenishment, budget release, and variable consumption driven by seasonal disease burdens—are common across pharmaceutical supply chains in many low- and middle-income countries (LMICs). As such, the insights and methodology developed here are not unique to Ethiopia but reflect broader systemic patterns typical of resource-constrained public health logistics environments. Also, the core contribution of this study lies not only in the specific predictors identified, but in the structured and replicable process we followed to elicit, validate, and integrate expert knowledge into forecasting models. This process—which includes expert engagement, anomaly detection through time series visualization, triangulation with operational records, and consensus-based feature engineering—can be readily applied in other settings, regardless of country-specific differences. To support broader applicability, we have made the forecasting framework and R and Python code publicly available. This allows practitioners in other countries to adapt the approach by substituting context-specific predictors while

retaining the general workflow. For example, countries with different inventory schedules or seasonal disease profiles can customize the model inputs accordingly, while benefiting from the same expert-informed, operationally grounded forecasting strategy. We agree with the reviewer that it is important to identify transferable methodological principles, and we believe our work contributes to this goal by offering a transparent and adaptable approach that balances local specificity with global relevance. Additionally, we have also added a new subsection that provides managerial implications that might be useful for other LMICs :

This study offers actionable lessons for supply chain managers, public health planners, and policymakers navigating the uncertainty of pharmaceutical demand—particularly in systems like Ethiopia's, where operational constraints and demand volatility are the norm, not the exception.

At present, national forecasting at the EPSS still relies heavily on basic extrapolation tools—usually Excel sheets or donor-developed software such as the Quantification Analysis Tool (QAT). These tools are easy to use but limited: they assume stable demand, ignore uncertainty, and often miss the operational signals embedded in local experience. In practice, forecasts are sometimes adjusted based on gut feeling or anecdotal program insights—not because planners want to—but because the tools don't offer a better alternative.

The models developed here—freely available, open-source, and designed to integrate directly with routine EPSS data—allow planners to make decisions using full probabilistic forecasts, not just single-point estimates. Forecasts are delivered with confidence intervals (e.g., 80% or 90%) that help teams decide not only how much to order, but how much risk they're willing to tolerate. For products with known seasonality—like antimalarials or diagnostic kits—this matters. The system doesn't just forecast a number; it gives planners a buffer strategy.

Beyond accuracy, the models also reflect how medicines actually move through the system. Predictors based on warehouse replenishments, fiscal year inventory routines, and known disease cycles are embedded into the forecasts—not hard-coded, but learned from the data in ways that are transparent and reproducible. These inputs are drawn from expert operational knowledge and can be updated or modified as the system evolves.

Importantly, the entire modeling pipeline is built in R and Python, and is already shared alongside code and data. This makes it immediately usable by EPSS analysts and regional logistics teams, even in settings without access to proprietary tools or high-end infrastructure. It's not a black box; it's something the system can own, modify, and grow with.

To support real uptake, we recommend five practical steps:

- Embed these probabilistic models into EPSS quantification rounds to replace rigid extrapolation methods and bring scenario-based planning into monthly and quarterly cycles.
- Use prediction intervals to define risk-informed buffer policies, especially for high-volatility products. Not every item needs the same margin of error.
- Automate the collection of operational predictors (e.g., replenishment events, stock counts, seasonal triggers) into the logistics data stream to reduce reliance on manual elicitation while preserving contextual intelligence.
- Train regional and central teams using the open-source tools and shared codebase, turning forecasting into a hands-on, in-country competency rather than a dependency

on external tools or consultants.

- Establish routine forecast review sessions, supported by visual dashboards that reflect not just "what the model says," but how uncertain that estimate is—and what that means for stock levels, procurement plans, and emergency readiness.

Taken together, these steps push forecasting beyond formulas and into real decision-making.

Comment 3: Regarding the model performance, the study compares multiple forecasting models (e.g., ARIMA, LSTM, TimeGPT), but it does not provide sufficient analysis of why certain models outperform others. For instance, the underperformance of LSTM models is attributed ambiguously to data characteristics, without profounder investigation into the specific limitations of the model or dataset. To tackle this point, the authors must conduct a thorough analysis to identify why advanced models like LSTM fail to perform as expected. Besides, they should compare models not only on accuracy but also with other relevant KPIs like computational efficiency, scalability, or utility.

Response: Thank you for this thoughtful and constructive comment. We fully agree that a more comprehensive analysis of utility, scalability, and computational efficiency would provide a valuable complement to the forecasting performance results.

In particular, we acknowledge the importance of assessing computational feasibility, particularly in resource-constrained environments such as Ethiopia. Understanding the runtime and complexity of different models is essential to inform decisions about practical deployment and sustainability in low-resource operational settings. In response to your suggestion, we have expanded our analysis to include a runtime comparison across all forecasting models used in the study. This analysis quantifies the total time required for model training and forecast generation across the full panel of 33 pharmaceutical products. The results of this comparison are now presented and discussed in a newly added subsection of the manuscript.

At the same time, we would like to clarify that the primary objective of this study was to examine the extent to which expert-informed contextual variables can enhance forecast accuracy in pharmaceutical supply chains. Therefore, our primary focus was on model accuracy and methodological rigor, with particular attention to the structured integration of domain knowledge into the forecasting pipeline. That said, we fully agree that future work should move beyond pure forecasting performance to evaluate utility in operational terms—such as how forecasts are used to inform procurement planning or inventory replenishment decisions. We see immense value in extending this line of research in collaboration with EPSS to explore how forecasting outputs can be systematically integrated into decision-making processes and to define appropriate utility metrics for evaluating real-world impact. We intend to pursue this in subsequent studies, building on the methodological foundation established here.

Moreover, we have expanded the discussion of the LSTM model's performance, with a more detailed analysis now included in the Results and Discussion section (Section 4).

We have introduced a new subsection within Section 4 to highlight the computational requirements of the various forecasting models:

Computational efficiency and resource considerations

In addition to forecast accuracy, it is also important to consider the computational efficiency of each model, particularly in settings where computational resources and technical expertise are limited.

Table 3 shows the total runtime required to train and generate forecasts across all 33 pharmaceutical product time series for each method for all origins. All models were implemented using R, except TimeGPT, which was run via Google Colab using a T4 GPU backend. All other models were executed on a local machine with 7 CPU cores (11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40 GHz and 8 GB RAM).

Table 3: Computation time required for training and generating forecasts for each model across all products.

Model	Runtime (Seconds)	Runtime type	category
LSTM	6892.30	CPU with 7 cores	high
LSTM with regressors	7234.00	CPU with 7 cores	high
Regression	27.48	CPU with 7 cores	low
Regression with regressors	47.38	CPU with 7 cores	low
ARIMA	96.27	CPU with 7 cores	low
ARIMA with regressors	138.08	CPU with 7 cores	medium
TimeGPT	2.57	Colab T4 GPU	low
TimeGPT with regressors	3.72	Colab T4 GPU	low
ETS	75.71	CPU with 7 cores	low

As shown in Figure 4 and 5, models that incorporated expert-informed regressors generally achieved better forecast accuracy compared to their univariate versions. This trend was most evident in classical models such as ARIMA and regression, where the inclusion of regressors led to a noticeable shift toward lower and more concentrated error distributions. However, this improvement came with an increase in computational cost. In all cases, the addition of regressors increased runtime—by 72% in the regression model (from 27.48 to 47.38 seconds) and by 43% in ARIMA (from 96.27 to 138.08 seconds). TimeGPT also showed a slight increase in runtime (from 2.57 to 3.72 seconds), although the total processing time remained extremely low overall. Moreover, LSTM exhibited the largest increase in runtime when regressors were added, rising from 6,892 to 7,234 seconds. This sharp increase reflects the sensitivity of neural networks to input configuration, especially in contexts with limited data, irregular demand, and noisy signals.

These results underscore the importance of balancing model performance with implementation cost. While deep learning models such as LSTM offer strong potential when well-tuned, they demand significantly more computational resources and may be less robust when integrating static or weakly aligned contextual features. By contrast, TimeGPT, a foundational model pretrained on large-scale time series data, provides a compelling alternative. It required less than 4 seconds to forecast all products, offered competitive accuracy, and required no tuning or retraining—making it well-suited for practical use in low-resource environments.

To further elaborate on the performance of the LSTM model, we have added the following section.

While LSTM models achieved the best overall forecast accuracy across products, their

performance exhibited notable variability depending on the characteristics of individual demand patterns. For example, the product "Amlodipine - 5mg - Tablet" demonstrates periods of extreme variability, with spikes in demand followed by periods of very low or zero consumption. Such patterns align well with the strengths of univariate LSTM models, which are adept at capturing long-term dependencies and managing complex temporal fluctuations. In contrast, the demand for "Anti-Rho (D)" is erratic and sparse, with frequent random fluctuations and little structural consistency. This lack of clear temporal patterns can make it challenging for LSTM models to learn generalizable signals, particularly given the limited data length available for training. Although LSTM can manage irregular data to some extent, it performs best when patterns are consistent or cyclic. These variations across product types contributed to the observed distribution of forecast errors across the product portfolio. Moreover, although we expect that multivariate LSTM models benefit from expert-informed predictors, our results show that the univariate LSTM consistently achieved better forecast performance. This outcome may be attributed to the nature of the predictors—binary, static, or weakly aligned with short-term temporal dynamics—which can disrupt rather than enhance learning when added to a neural network sensitive to input configuration. Moreover, with relatively short historical series and limited training samples, the inclusion of additional variables may have led to overfitting or reduced generalization. These findings suggest that while LSTM models can effectively learn temporal patterns from consumption data alone, incorporating structured external knowledge requires careful feature engineering and alignment to be beneficial.

Comment 4: Certain methodological aspects lack clarity. For example, the process for selecting and validating predictors derived from expert insights is not adequately detailed and also the description of model training and hyper parameter tuning lacks explanations. Therefore, providing a detailed explanation of how predictors were identified, validated, and integrated into the models is valuable, and, including specifics on hyper parameter optimization, training processes, and cross-validation to enhance methodological transparency.

Response: Thank you. To systematically translate domain knowledge into the forecasting models, we used a structured approach including expert identification, consultation, exploratory data analysis, and the formal translation of insights into model-ready variables. In the paper, each step of this process is outlined in detail to address the reviewer's concerns and to illustrate how the approach can be adapted to other contexts. We also refer the reviewer to our response to your Comment 1, where we provided a comprehensive explanation of how relevant contextual factors were identified in collaboration with domain experts and subsequently encoded as deterministic predictors.

We agree that we didn't explain the hyperparameter tuning and describe the cross- validation. Now we have added the following section. Now, we have expanded the experiment setup section to include a detailed description of our time series cross-validation approach. Specifically, we adopted a rolling-origin expanding window strategy with a 6-month forecast horizon and monthly rolling updates, aligned with EPSS planning requirements. We also clarified that model development and hyperparameter tuning were conducted strictly within the training data at each iteration to maintain evaluation integrity. Now we have added the following section.

The following section was added to the Long Short-Term Memory neural network (LSTM) in section 4:

To improve the robustness of the LSTM models and address overfitting issues, we introduced dropout regularization layers after the LSTM units and employed early stopping based on validation loss during model training. Each LSTM model was trained independently for each product series. Hyperparameter tuning was performed in a preliminary phase using a subset of products. Various configurations of LSTM units (30–100), dense units (50–200), dropout rates (0.1–0.5), and batch sizes (16–64) were evaluated based on validation set performance. The final model structure — 50 LSTM units, 100 dense units, a 0.2 dropout rate, and a batch size of 32 — was selected as a trade-off between forecast accuracy and model stability across different demand patterns.

We have described the cross validation in the revised manuscript:

To evaluate the performance of our forecasting models, we employed a time series cross-validation approach, following best practices for forecasting evaluation (Hyndman and Athanasopoulos, 2021). Rather than using a fixed training and test split, we used a rolling-origin cross-validation framework, which allows for a more comprehensive assessment across different demand patterns and periods.

In our setup, the initial training set consisted of all available historical data from December 2017 up to June 2021. The evaluation period covered the subsequent 12 months, reflecting the operational needs of the EPSS, which plans consumption over a one-year horizon. At each iteration, models were trained on an expanding training window and evaluated over a fixed 6-month forecast horizon, aligned with typical EPSS planning cycles. After each forecast generation, the training set was expanded by one additional month, and the process was repeated, allowing for rolling assessment across the final 12 months of data.

This structure ensured that forecasts reflected realistic operational scenarios, where forecasts are continuously updated as new data becomes available. Model development and hyperparameter tuning were strictly confined to the training sets at each origin to prevent information leakage. For probabilistic evaluation, 1,000 future paths were simulated per series, enabling robust estimation of forecast uncertainty.

This cross-validation design allowed us to evaluate each model's ability to perform across multiple different forecast origins and a variety of demand conditions, providing a more reliable and generalizable understanding of model performance.

Furthermore, all model development steps, including hyperparameter tuning for the LSTM models, were conducted exclusively using the training data available at each iteration. No test set information was used during model selection or tuning.

Comment 4: Furthermore, the study represents an insufficient probabilistic forecasting insight. In fact, metrics like CRPS are presented, but their practical implications are not adequately explored. It is mandatory to provide practical examples or case studies to illustrate how probabilistic forecasts can enhance operational planning and decision-making in pharmaceutical supply chains.

Response 4: Thank you. We have added a paragraph using an illustrative case as requested by another reviewer to discuss how forecast might be useful including probabilistic forecast:

To further illustrate the practical relevance, we consider a case where the mean forecasted consumption for the next month is 1,000 units, with a 90% prediction interval ranging from 850 to 1,150 units.

Under a traditional approach, inventory decisions would rely solely on the point forecast. Based on the mean prediction of 1,000 units, the store manager would order precisely that quantity, assuming it would meet the expected demand. However, this approach does not incorporate any adjustment for uncertainty, potentially leading to stockouts if actual demand exceeds 1,000 units, or excess inventory if consumption is lower.

In contrast, utilizing the full probabilistic forecast allows the decision-maker to take variability into account. If a higher service level is required, for example 95%, the inventory policy may be adjusted by stocking closer to the upper bound of the 90% prediction interval (e.g., 1,150 units). If inventory holding costs are a major concern and the organization can tolerate a modest risk of shortage, the order quantity could remain closer to the median forecast of 1,000 units.

Table 1: an illustrative example of using point and probabilistic forecasts

Forecast Type	Ordering Quantity	Risk Consideration
Point Forecast Only	1,000 units	No explicit consideration of uncertainty
Probabilistic Forecast (95% service level)	1,150 units	Adjusts inventory to account for demand variability

This illustrative comparison indicates that point forecasts provide a single estimate without adjusting for forecast uncertainty, while probabilistic forecasts allow decision-makers to explicitly align inventory decisions with risk tolerance and service level requirements. This is particularly important for pharmaceutical products, where the consequences of stockouts or overstocking can be significant both operationally and clinically.

The following is also highlighted in the paper:

In practice, point forecasts are commonly used despite their limitations, but they do not account for the inherent uncertainty associated with forecasts. The future is inherently uncertain, and effective planning requires considering alternative scenarios. Probabilistic forecasts offer a comprehensive approach by assigning likelihoods to a range of possible outcomes, recognizing that different consumption levels may occur with varying probabilities. The goal is to maximize the sharpness of these predictive distributions—i.e., how concentrated the forecasts are—while maintaining calibration, meaning that the predicted probabilities align well with actual outcomes (Gneiting and Katzfuss, 2014). This approach allows decision-makers to fully leverage available information, incorporating uncertainty in a structured and measurable way. The primary purpose, as illustrated in Figure 6, is to quantify and communicate uncertainty. This figure displays the forecast distribution of consumption over a 6-month horizon using a density plot. For each month within the forecast period, a separate distribution is generated. The plot also includes the point forecast alongside 80% and 90% prediction intervals to illustrate potential variability.

Comment 5: Moreover, the practical implications should be improved. The study lacks actionable recommendations for stakeholders in pharmaceutical supply chains. The authors identify key challenges, such as erratic demand and resource constraints, but they do not offer

concrete solutions for addressing these issues. In other words, the authors do not propose actionable steps for integrating the forecasting methodology into existing supply chain systems or highlight how policymakers and practitioners can use the findings to improve supply chain management, and decision-making.

Response: Thank you for this valuable comment. We agree that providing actionable recommendations would enhance the practical relevance of our findings. In response, we have revised the discussion section to include concrete suggestions for integrating the forecasting framework into existing pharmaceutical supply chain systems. We have also outlined how policymakers and practitioners can use the insights from our study to support data-driven decision-making and improve resource planning, particularly in low-resource settings like Ethiopia.

Managerial Implications

This study offers actionable lessons for supply chain managers, public health planners, and policymakers navigating the uncertainty of pharmaceutical demand—particularly in systems like Ethiopia's, where operational constraints and demand volatility are the norm, not the exception.

At present, national forecasting at the Ethiopian Pharmaceutical Supply Service (EPSS) still relies heavily on basic extrapolation tools—usually Excel sheets or donor-developed software such as the Quantification Analysis Tool (QAT). These tools are easy to use but limited: they assume stable demand, ignore uncertainty, and often miss the operational signals embedded in local experience. In practice, forecasts are sometimes adjusted based on gut feeling or anecdotal program insights—not because planners want to—but because the tools don't offer a better alternative.

The models developed here—freely available, open-source, and designed to integrate directly with routine EPSS data—allow planners to make decisions using full probabilistic forecasts, not just single-point estimates. Forecasts are delivered with prediction intervals (e.g., 80% or 90%) that help teams decide not only how much to order, but how much risk they're willing to tolerate. For products with known seasonality—like antimalarials or diagnostic kits—this matters. The system doesn't just forecast a number; it gives planners a buffer strategy.

Beyond accuracy, the models also reflect how medicines actually move through the system. Predictors based on warehouse replenishments, fiscal year inventory routines, and known disease cycles are embedded into the forecasts—not hard-coded, but learned from the data in ways that are transparent and reproducible. These inputs are drawn from expert operational knowledge and can be updated or modified as the system evolves.

Importantly, the entire modeling pipeline is built in R and Python, and is already shared alongside code and data. This makes it immediately usable by EPSS analysts and regional logistics teams, even in settings without access to proprietary tools or high-end infrastructure. It's not a black box; it's something the system can own, modify, and grow with.

To support real uptake, we recommend five practical steps:

- Embed these probabilistic models into EPSS quantification rounds to replace rigid extrapolation methods and bring scenario-based planning into monthly and quarterly cycles.

- Use prediction intervals to define risk-informed buffer policies, especially for high-volatility products. Not every item needs the same margin of error.
- Automate the collection of operational predictors (e.g., replenishment events, stock counts, seasonal triggers) into the logistics data stream to reduce reliance on manual elicitation while preserving contextual intelligence.
- Train regional and central teams using the open-source tools and shared codebase, turning forecasting into a hands-on, in-country competency rather than a dependency on external tools or consultants.
- Establish routine forecast review sessions, supported by visual dashboards that reflect not just "what the model says," but how uncertain that estimate is—and what that means for stock levels, procurement plans, and emergency readiness.

Taken together, these steps push forecasting beyond formulas and into real decision-making.

Comment 6: The study highlights also some significant challenges such as disruptions from COVID 19 and conflicts, though these are not analysed in depth. These disruptions introduce inconsistencies that complicate modelling and forecasting, yet the manuscript offers little in terms of solutions. The authors should analyse the impact of these disruptions on forecasting accuracy and model performance and propose methods for accounting for such disruptions, such as scenario-based forecasting or adaptive models. Overall, the paper addresses an important topic and demonstrates potential contribution, but it requires substantial revisions to address the previous identified concerns.

Response: Thank you for this important observation. We fully agree that addressing the impact of disruptions such as COVID-19 and armed conflicts is critical, particularly given their potential to introduce irregularities in demand patterns and to complicate both model training and evaluation. However, isolating and quantifying the effects of such disruptions is far from straightforward. Although some experts at the workshop highlighted these factors, there was no consensus that they had a measurable impact on the consumption of the 33 products analyzed in this study.

Nonetheless, we believe that further research is needed to better understand, and model disrupted demand, particularly within the context of pharmaceutical supply chains in low- and middle-income countries (LMICs). This includes addressing issues such as censored or unmet demand, as well as the effects of epidemics and other structural disruptions like conflicts. Several methodological approaches have been proposed to handle such structural breaks, including those recently discussed by Hyndman and Rostami-Tabar (2025). These approaches span scenario-based forecasting, treating the disruption as a missing data period, model adaptation, and intervention analysis. We also believe there is significant potential in developing causal forecasting models that explicitly account for country-level policies, epidemic dynamics, and socio-economic factors in addition to historical consumption patterns. This represents a promising direction for future research. We have included this in future directions of research.

We have added the following text in the revised manuscript:

Disruptions such as the COVID-19 pandemic and armed conflicts may alter consumption patterns in time series data, introducing irregularities that challenge the assumptions of traditional forecasting methods. In such contexts, it becomes important not only to model

observed consumption but also to account for censored demand—i.e., the unmet need that would have been fulfilled under normal conditions and is essential for effective replenishment planning. Future research could focus on developing and evaluating forecasting strategies that are specifically designed to handle disrupted time series and estimate censored demand. In particular, methods that enhance the robustness of probabilistic forecasts in the presence of structural breaks or exogenous shocks are especially relevant. One promising direction is also the development of causal forecasting models that incorporate policy changes, epidemic dynamics, and socio-economic factors—an approach that may be particularly well suited to the realities of pharmaceutical supply chains in low- and middle-income countries (LMICs).