"HOW LIKELY IS AN INFLATION DISASTER? A QVAR-BASED STUDY OF KAZAKHSTAN".

BY
AKBOTA BOKANOVA

THESIS

Submitted in partial fulfillment of the requirements

for the degree of Master of Science in Finance

in the Graduate School of Business

Nazarbayev University, 2025

Astana, Kazakhstan

Advisor: Dr. David De Remer

ABSTRACT

This research question examines the tailing of inflation in Kazakhstan using a Quantile Vector Autoregression (QVAR) framework to monthly macroeconomic data between 2005-2024. Nonlinear and state-dependent inflation in a small open economy as a commodity exporter cannot be described using standard VAR and BVAR models where assumptions include symmetric disturbances and are Gaussian. Based on the QVAR methodology of the ECB, the research measures quantile-specific transmission, quantile impulse responses, and Inflation-at-Risk (IaR). Findings indicate that despite the prevalence of exchange-rate shocks in general mean inflation dynamics, tail outcomes are influenced mainly by inflation persistence with the influence of external shocks being less strong once the economy has entered the high-inflation regimes. The stress-testing exercises, such as artificial shocks, ECB-type stress testing over several months, show that extreme inflation risks increase exponentially when subjected to long-term depreciation, supply shocks or price inertia. Lastly, out of sample forecasts illustrate that QVAR works far better than VAR and random-walk forecasts at upper-tail inflation forecasting, with smaller pinball loss and accurate quantile coverage. In general, the thesis presents the initial QVAR-based assessment of Inflation-at-Risk that applies to Kazakhstan and the distribution-conscious framework that can be effectively used in monetary policies and risk monitoring.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. David De Remer, whose guidance, support, and patience made this thesis possible. Throughout this process, he has been exceptionally generous with his time, always willing to discuss ideas, clarify concepts, and help me navigate challenges both technical and conceptual. His thoughtful feedback pushed me to think more critically, refine my arguments, and improve the quality of my work at every stage.

I am truly grateful for his mentorship-not only for the expertise he shared, but also for the encouragement and confidence he continuously offered. Working under his supervision has been an inspiring and transformative experience, and I sincerely appreciate the dedication he demonstrated towards my academic growth.

INTRODUCTION

External exposure, structural change of regime, and high global uncertainty is causing the dynamic nature of inflation in emerging economies to be more challenging to unravel. The state-dependent, nonlinear, fat-tailed nature of inflation is beyond the ability of standard models based on the mean. Kazakhstan is a good example of this difficulty: being a small open commodity exporter, high exposure to the exchange-rate pass-through, and periodic changes in the monetary policy, Kazakhstan has significant inflation tail risks, which are systematically underpriced by conventional forecasting methods.

Adrian, Boyarchenko, and Giannone (2019) proved that risks relevant to the policy do not cluster around the mean, and this has led to the introduction of Growth-at-Risk framework, which is subsequently taken up by IMF and ECB. This was extended to inflation by recent literature. Banerjee et al. (2024) identified very asymmetric distributions of inflation where tail risks accelerate during commodity and financial shocks and Makabe and Norimasa (2022) established tail risks depending on forecast horizons. These analyses verified that quantile techniques can help identify weaknesses that the conventional modeling does not offer.

This is a vital question as far as monetary policy in emerging markets is concerned. The traditional VAR models with the assumption of the symmetric response systematically fail to reflect the severity of the inflation surge during crisis episodes, e.g. the 2015 currency crisis in Kazakhstan when the inflation rates went up by 4% to 14% or the commodity shock of 2022 exceeding 20%. During this time the central banks were given overly optimistic forecasts at the time when risk assessment is most needed. When predicting manageable average inflation is the goal of inflation-targeting economies such as Kazakhstan, average tail risks being extreme in fact, changes the choices made by the policies, particularly whether to preempt and aggressively

tighten or wait until inflation occurs, at which point it may also be the case that expectations are now de-anchored.

This study is crucial because no research has systematically evaluated the Inflation-at-Risk of Kazakhstan with Quantile Vector Autoregression (QVAR) -the ECB tool of Chavleishvili and Manganelli (2020). Current literature is based on VAR or BVAR models which make assumptions of symmetric disturbances and cannot explain extreme results and thereby compromise risk-oriented monetary policy.

This research has three contributions. First, it redefines the QVAR structure based on monthly data (2005-2024) to adapt to the macro-financial structure of Kazakhstan using the identification characteristics of an oil-exporting economy. Second, it achieves quantile specific impulse responses and ECB-type multi-month stress testing in order to measure the variation between the behavior of inflation tails and median behavior during shock conditions. Third, it also compares the performance of QVAR forecasting to that of VAR and random walk performance using rolling out of sample tests.

Moreover, the practical value for the National Bank of Kazakhstan (NBK) and similar emerging market central banks becomes clear through three policy applications.

First, the stress-testing system can facilitate analysis with scenarios that change the nature of policy questions to be asked rather than what will average inflation be, but what is the worst-case scenario and what is its probability of occurrence, which is important when the cost of breach of the target band is credible. The ability to measure the extent to which sustained exchange-rate pressures or commodity shocks reinforce into radical inflation consequences over several months

allows policymakers to defend preemptive tightening using tail-risk probabilities instead of allowing median forecasts to fail.

Second, the methodology offers the capacity of early-warning on the accumulation of inflation risks in the upper tail that is asymmmetric. When inflation is in high-risk regimes, the process of persistence starts to take over and the transmission of a monetary policy becomes weak. Early identification of such a transition can be translated into intervention at a time when policy instruments work well enough, much more effectively than responding to inflation spikes. Surveillance of the occurrence of upper-tail risks in disproportionate relation to the median predictions assists in initiating precautionary steps that allow the avoidance of entry into the persistence dominated regimes.

Third, the model augments the awareness on the inflation behavior upon the actualization of these high-risk regimes. The QVAR outcomes show that the inflation dynamics at the upper tail become more self-reinforcing and thus the lagged inflation influences more than current shocks. This directly concerns policy importance: getting price stability back is much more expensive when the persistence is increased. The framework can therefore provide policymakers with a better understanding of how fast or slow inflation can be by mapping these nonlinear regime dynamics empirically with regard to different starting conditions.

LITERATURE REVIEW

The idea that inflation should be examined not only in the context of its average behavior but also in the context of the tail risks that arise during times of economic stress is beginning to take center stage in the present developments of macroeconomic studies.

A growing body of evidence shows that inflation is not symmetric, various negative shocks are more likely to cause inflation to overburden the high-tail than to cause balanced disinflationary movements. Queyranne M., Lafarguette R, and Johnson K. (2022) in their “Inflation-at-Risk in the Middle East and Central Asia” state that most countries, including Kazakhstan, experienced increase in the inflation asymmetry overtime. They found that overtime inflation distribution became positively skewed, the kurtosis, which measures how fat the tails of the inflation distributions are compared to a normal distribution, has increased along with the variance. Huang & Zhang (2023) and Banerjee et al. (2024) named oil price shocks, monetary policy shocks, exchange rate pressures and inflation expectations as main shocks that pushes inflation disproportionately toward its upper tail, causing these assymetries. These findings show that inflation tail risks have become more permanent and traditional mean-based approaches as classical VARs are not well-suited for understanding the full dynamics of inflation anymore. This has motivated the shift toward Inflation-at-Risk frameworks, which concentrate on the probability of severe and economically significant tail events and explicitly represent the complete conditional distribution of inflation. Instead of asking how inflation responds “on average,” the IaR framework evaluates how different shocks influence the worst-case inflation scenarios, which are often the most relevant for emerging economies exposed to commodity volatility, exchange-rate pressures, and policy regime shifts. The IaR concept is naturally complemented by quantile regression and quantile VAR approaches, which capture nonlinearities and state-dependent behavior by allowing the response of inflation to vary across different points of its distribution. Unlike standard VARs, which aim at estimating only mean responses, this model analyzes how shocks propagate during periods of low, median, or high inflation, highlighting whether certain shocks disproportionately affect the upper tail. The analysis of

Inflation-at-Risk by the ECB (ECB Working Paper No. 2330) shows that such models are especially useful in modelling nonlinear, asymmetric, and regime-dependent inflation, thus effective for analyzis of economies such as Kazakhstan's, where the behaviour of inflation can change suddenly in response to external and internal shocks.

Chavleishvili and Manganelli (2020) presented the methodological background of dynamic tail-risk modeling in the Working Paper No. 2330 by European Central Bank, titled Forecasting and Stress Testing with Quantile Vector Autoregression. Their contribution generalized the standard VAR model to the quantile so that terms are allowed to interact between variables anywhere in the quantile space of their joint distribution. This innovation could detect the behavior of the economy in the stressful periods, when financial shocks cause the system to jump into the lower or upper tail of the possible outcomes. The authors used a recursive decomposition of the joint distribution of macro variables in their Quantile VAR (QVAR) model and demonstrated a series of successive conditioning of the multivariate quantiles by the application of their law of iterated quantiles. This law allows the future quantiles to be predicted recursively. They also came up with the idea of a quantile selection matrix that determines the paths along which certain quantiles occur over time. By feeding a series of tail quantile shocks into the system, this structure enabled the creation of stress scenarios like six months of straight shocks on the financial side.They showed the application of a recursive Cholesky-type ordering to structural identification with the help of QVAR, which in turn allowed computation of Quantile Impulse Response Functions (QIRFs) – responses of any element of the distribution to structural shocks.

The highly asymmetric whichmpact of financial shocks on the left end of production growth was found in their empirical application to the euro area, where they relied on industrial production

and a financial stress index (CISS): the financial shocks had a big and long-lasting effect on the left tail of production growth, whereas the impact on the average effects was weaker with standard VARs. In addition, they engaged in stress-testing exercises on central bank scenario analysis by designing “tail quantile sequences”. For example, six-month series of pessimistic (10 th -percentile) financial shocks would lead to four percentage points reduction in output in 2008, which is twice as much as a median forecast would indicate. This theoretical framework brought together three goals methodologically:

First, quantile forecasting-prediction of macro variables when states change.

Second is stress testing- the construction of series of quantile-realizations of adverse algorithms.

Third is structural interpretation-the asymmetric transmission of the shock in the quantiles.

These characteristics made QVAR a significant improvement on the direct quantile regressions (Adrian et al., 2019) and nonlinear mean-based models (Kilian and Vigfusson, 2017). Hence, since 2020 QVAR and similar quantile techniques have been used on a variety of macro-financial subjects. Banbura, Hubrich, and Lenza (2024) built on the approach to Quantile Structural VARs (Q-SVARs), which includes quantile dynamics and structural identification constraints to policy stress testing. Based on time-varying quantile regressions, Korobilis et al. (2021) found out the dynamics of the inflation risks in the euro area to illustrate that tail behaviors vary according to the monetary regimes and international shocks. Schule (2020) utilised a quantile structural VAR to investigate asymmetric effects of uncertainty on the business cycle whereas Montagnoli and Napolitano (2022) a QVAR to oil price shocks and inflation asymmetries, and the oil-inflation channels that resource-dependent economies such as Kazakhstan are primarily keyed to. In the meantime, in another example of the emerging market, the IMF in ““Inflation-at-Risk” in the

Middle East, North Africa, and Central Asia" (2022) used quantile regressions of inflation risk, marking that the upper-tail inflation predominantly occurs as a result of commodity and exchange-rate shocks. Ahmed et al. (2024) also demonstrated that the use of inflation-targeting regimes considerably lowers inflation tail risk, particularly in unstable economies - a fact that is directly applicable to the post-2020 monetary reforms to be implemented in Kazakhstan.

Collectively, these papers point out that changes in inflation tail risk are state-specific as they are intense when there are stress shifts globally and more modest when monetary credibility is plentiful. They also highlight that quantile-based VAR and regression techniques are emerging as the centre of macro-financial policy making and risk supervision.

Based on this body of literature, the current paper applies the QVAR framework of Chavleishvili and Manganelli (2020) to a developing economy and exporter of commodities, such as Kazakhstan, where the volatility of inflation is conditioned by oil prices, exchange rate processes, and monetary policy reactions. Although earlier applications have been done on the euro area or on global samples, country-level Inflation-at-Risk modeling of resource-dependent and small open economies has a gap.

The contribution of this study is threefold:

1) Methodological Adaptation: adapts the Quantile Vector Autoregression strategy to the data structure and macro-financial transmission mechanisms in Kazakhstan, based on a recursive identification order (oil - FX - policy rate - output - inflation), appropriate to an oil-exporting economy. This model includes monthly data (2005-2024), quantile impulse responses, and multi-month stress simulation similar to ECB stress tests.

2) Regime Analysis: In contrast to earlier applications of QVAR, this paper does take into consideration monetary regime shifts, including most prominently a 2020 shift toward a stronger inflation-targeting regime and flexible exchange rates. The findings indicate an exchange-rate pass-through decrease of 75 percent after 2020, which indicates the way in which institutional credibility remakes inflation tail behavior. To evaluate model reliability, the study combines multi-month shock simulations, historical stress testing, and coverage backtesting. It concludes that although QVAR and VAR perform similarly for average inflation, QVAR is unique in its ability to capture clustered risks and tail asymmetries, particularly during crises like 2008, 2015, 2020 and 2022.

In conclusion, this paper shows that Kazakhstan's inflation risks are highly regime-dependent and externally driven by operationalizing the QVAR stress-testing framework in an emerging-market context. The first systematic estimation of inflation-at-risk for Kazakhstan is provided by integrating local structural insights with the ECB quantile methodology. This approach offers both methodological innovation and immediate policy relevance for the National Bank "inflation management strategy.

DATA AND DATA TRANSFORMATION

This section describes the construction of the monthly macroeconomic dataset used in this empirical analysis. The data processing consists of two stages- data cleaning, performed fully in Stata and data transformation, performed in R. The objective of this section is to develop monthly panel for Kazakhstan, covering the period from January 2005 to December 2024 and including information on inflation, industrial production, exchange rates, oil prices, interest rates, and inflation expectations, all of which are harmonized to a standard frequency.

Almost all raw data, except inflation expectations, was extracted from Bloomberg Terminal (Bloomberg L.P.), using standard Bloomberg tickers and Excel export functionality. These extracts were combined into 1 Excel and imported into Stata.

The cleaning process constructed from standardizing the raw date variables into unified monthly form, sorting them chronologically and dropping any duplicate months so that Stata could read the data. Additionally, months that did not have information on any of the explanatory variables were dropped. If orginially the datasets covered years from 1990 to 2025, after this cleaning process final sample was restricted to January 2005 to December 2024, ensuring complete availability across all variables.

The dependend variable- CPI inlfation rate *(pi_yoy)* was extracted in Year over Year (YoY) format, not as levels, as it helps to remove seasonality and measurement noise. Moreover, official inflation target of NBK is also expressed in YoY format. What comes to independent variables, industrial production index which measirs economic's real activity, was extracted in MoM terms. Similarly, this format reflects already seasonally adjusted indicators. The reason why this variable was included as one of the inflation drivers is that it links supply capacity to price pressures and strong domestic demand or weakening in production are expected to affect inflation.

Since Kazakhstan has import dependent economy with strong exchange rate pass through, depreciation in USD/KZT exchange rate leads to an increase in prices of imported consumer goods, fuel, and production inputs.

Kazakhstan is also one of major oil exporters, which means that oil revenue affects to country's budget expenditures, liquidity conditions and exchange rate dynamics. Therefore, third chosen

independent variable is Brent oil prices. It is expected that this indicator can influence inflation through demand and income channels.

Then, monthly NBK base rates were extracted from Bloomberg. It was crucial to add this variable to the model because monetary tightening raises funding costs and suppresses inflation, while monetary easing increases liquidity and stimulates demand. Hence, policy ineterst rates are essential to capture how monetary authorities respond to macroeconomic shocks.

Finally, inflation expectations of population were also included to the model as a forward looking component. It was extracted from NBK's household inflation expectations survey capturing February 2016 to 2025. It helps to evaluate anchoring of inflation under NBK's inflation targeting regime. Expectations influence consumer price settings, wage negotiations and spending decisions, hence it can be expected that inflation is partly driven by population's expectations.

Since Var/QVAR models require stationary variables, several data transformation steps were taken. Firstly, as exchange rate levels are non-stationary in its nature, they were converted into monthly log change (*dlog_fx*) to capture actual depreciation. ($100 \times \Delta\ln(\text{USD/KZT})$) Then, because oil prices fluctuate strongly over time, log change was computed to capture any short term shocks. ($100 \times \Delta\ln(\text{Brent})$). In order to ensure stationarity for other key variables too, differences of these variables were taken. Finally, as oil price and exchange-rate log-changes had extreme outliers,winsorization at the 1–99% level was applied. In QVAR models which are highly concentrated around tail values, such large fluctuations can dominate the whole distribution which consuquently distort the computed quantiles and lead to unstable coefficients. Winsorizing in such cases prevents influence of extreme shocks on the results while still preserving the overall pattern of the data.

After all variables were transformed and data sample was clean to use, to ensure that variables used in the model satisfy the assumptions required for VAR-based estimation, all transformed macroeconomic indicators were subjected to Augmented Dickey–Fuller (ADF) unit-root testing. According to the results of these testing, all variables have p-values smaller than 0.01, which confirms that series are suitable for VAR and QVAR estimations.

After stationarity test was passed, the dataset was examined for a possible structural break associated with Kazakhstan's transition to a floating exchange-rate regime in August 2015. For this purpose, Chow breakpoint test was applied to the inflation equation, with the break point set to January 2016, because the IMF's Annual Report and classification lists this date as the official start of the floating regime, as only by early 2016 had the NBK stopped interventions fully. Results of the test was considered as statistically significant with p-value equalling 0.039 at the 5% level. This indicates that the relationship between inflation, exchange-rate changes, oil-price shocks, and monetary policy shifted after the regime transition, supporting the empirical decision to explicitly account for post-2016 dynamics in the modelling framework. That is why along with the full sample, post-2016 period was examined separately.

METHODOLOGY

This chapter presents the methodological framework used to estimate the impact of exchange rate, oil price, real economic activity, and monetary policy shocks on the upper tail of Kazakhstan's inflation distribution. The analysis has been conducted according to Chavleishvili and Manganelli (2019, ECB Working Paper No. 2330) and is based on structural quantile vector autoregressions (QVAR), multi-step quantile impulse responses and ECB-style multi-month tail risk stress tests. The methodological contribution of the thesis is the adaptation of this framework to an emerging commodity-exporting economy of Kazakhstan where structural regime changes

and externally driven shocks took place, making inflation more volatile and hence harder to predict.

Modeling Framework Overview

According to the official publications of National bank for forecasting economic activities and inflation patterns BVAR (Bayesian Vector Autoregression) model is being used. This model is considered to be advantageous compared to classical VAR model because it applies Bayesian shrinkage priors in parameter estimating. Such approach helps to reduce overfitting and stabilize the model, which in turn results in prodicung more accurate and robust inflation forecasts. Kazakhstan's macroeconomic indicators exhibiting high volatility and limited historical depth make BVAR model appropriate to employ to project the expected path of inflation. It assumes symmetric, thin-tailed and Gaussian disturbances. However, strongly skewed, fat tailed inflation dynamics of Kazakhstan combined with high exposure to external shocks makes these assumptions inconsistent with observed data, which leads to inaccurate forecasts of upper tail inflation risks. Another assumption of BVAR model is that impact of the shocks are the same at any point of the inflation distribution, which is contradictory, as in reality in the upper tail of the inflation distribution the impact of the shocks are stronger. This makes BVAR model unable to identify the likelihood and the size of the potential extreme inflation outcomes.

To address these limitations, this thesis proposes new approach- Quantile Autoregression (QVAR), which models not only average response of inflation to shocks but captures entire distribution of inflation, allowing shock transmission to change across quantiles of the distribution. As a result, this model is able to capture "tail risk" and to produce reliable measures of inflation-at-risk.

2. Vector Autoregression (VAR) and Structural VAR (SVAR)

At the beginning of the analysis, in order to obtain average behavior of inlfation, standard VAR and structural VAR (SVAR) models were built. Altough the ECB methodology (Chavleishvili & Manganelli, 2019) does not include a structural VAR model, in this thesis their framework is extended by estimating VAR and SVAR model via Cholesky decomposition in order to provide a clear benchmark of Kazakhstan's mean-based transmission mechanism before turning to quantile dynamics.

The VAR model provides a simple, data-driven description of how each variable changes over time in relation to its own lag and the lag of the other variables in the system. That is, it demonstrates how the oil prices, exchange rate, policy rate, real activity and inflation go together and affect each other on a monthly basis. The lag length was selected to be equal to 2 because monthly macroeconomic variables usually require lag length to be between 1 to 3 and for this dataset the optimal lag length was found to be equal to 2. This model is widely used by central banks such as the US Federal Reserve, European Central Bank, Bank of England and researchers to study how several macro variables interact over time.

The vector in this case looks like:

$$y_t = \begin{bmatrix} dlog\_oil\_w \\ dlog\_fx\_w \\ d\_policy\_rate \\ ipi\_growth \\ d\pi_{yoy} \end{bmatrix}$$

The model itself is:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + c + u_t,\ u_t \sim N(0, \Sigma_u),$$

where A_1, A_2- autoregressive coefficient matrices, c is the constant term and u_t is a vector of residuals.

One of the main limitations of classical VAR model in this analysis is that reduced form shocks u_t reflect the unexplained part of the model without showing what type of economic disturbance exactly is responsible for the movement. For example, a sharp jump in inflation might be caused by oil-price shock, exchange rate depreciation or change in domestic demand. As classical model fails to distinguish between different kinds of shocks, to analyze how inflation reacts to certain disturbances amd generate impulse responses reflecting economic shocks, the SVAR model was introduced. Structural VAR model is the extension of classical VAR model which imposes economically motivated restrictions that allow identifying structural shocks. To be able to distinguish between u_t and ε_t, a contemporaneous causal structure named Cholesky decomposition, one of the most widely used identification strategies, was imposed into this system. This identification method assumes a recursive ordering of variables. Under this aproach, the covariance matrix of the reduced form shocks is:

$$\Sigma_u = E[u_t u'_t],$$

Factorized as : $\Sigma_u = BB'$, where B is the lower triangular matrix. The structural shocks then take the form of:

$$\varepsilon_t = B^{-1} u_t$$

where:

u_t- reduced firm shocks,

B- contemporaneous impact matrix,

ε_t-structural shocks.

This equation separates the mixed shocks under VAR system into meaningful shocks under SVAR:

$$\varepsilon_t=(\varepsilon_t^{oil},\varepsilon_t^{fx},\varepsilon_t^{policy},\varepsilon_t^{ipi},\varepsilon_t^{\pi})'$$

Consequently, SVAR structural form looks like:

$$By_t=K_1y_{t-1}+ K_2y_{t-2} +c+\varepsilon_t,\ K=B*A$$

A Cholesky decomposition is a way to identify structural shocks in a VAR model. The key assumption of Cholesky decomposition is that cariables appearing earlier in the model are the most exogenous ones, which means that they are contemporaneously less affected by others and variables at the end of the ordering are the most endogenous. The ordering that is the most accurate and suitable for Kazakhstan's economy is as follow:

dlog_oil_w → dlog_fx_w → d_policy_rate → ipi_growth → d_pi_yoy

Oil price is treated as the most exogenous because it is largely determined on global markets. Exchange rate – even if responds quickly to global shocks, it is not influenced contemporaneously by domestic variables. Real activity measures in the form of industrial production index reflects domestic economic conditions that react to oil and FX movements with short delays. Central banks react to external and internal conditions by changing policy rates, but it cannot affect oil price or exchange rate whichmmediately. Finally, whichnflation whichs the most endogenous variable responding to every other variable.

So the economic idea behind this ordering is such that global shocks affect domestic economic activity and monetary policy, which in turn have influence to inflation.

After identifying correct Cholesky ordering, an Impuls Response Functions (IRF) were generated. IRFs show how one variable in the VAR system responds over time to a one standard deviation structural shock in another variable. It is computated by the formula below:

$$IRF(h)=\frac{\partial Y_{t+h}}{\partial \varepsilon_t}, \text{ for } h=0,1,2\ldots,12$$

These steps can be considered as diagnostic steps which help to confirm that accurate macroeconomic ordering and lag structure will be used in the QVAR.

Quantile Regression (QR)

After classical VAR model provided the baseline transmission mechanism under mean assumptions and found to be incompatible with Kazakhstan's inflation being fat-tailed, skewed and responding strongly to economic shocks during stress periods, Quantile regression model was introduced. This model eases several assumtions that classical VAR model requires. It does not require normality and allows variance of error to change across different distributions. Hence, one of the advantages of this model is that it allows capturing the assymetric behavior of inflation by estimating different parts of the conditional distribution, not just the mean. For example, it can estimate 90th percentile of inflation distribution, i.e what happens in the worst 10% of months when inflation spikes sharply. Even though standard QR model assumes observations' independence over time, the modern time series Quantile regression relaxes this limitation by relying on bootstrap methods to get valid standard errors, confidence intervals and inference. This simple univariate quntile regressions helps to check how FX, oil, policy rate and domestic activity affect inflation at different quantiles. It provides a static, one period ahead estimate of the conditional quantile of monthly inflation $d\pi_t$.

The starting point of this regression model whichs contructing working dataset, whichh contains the date, regime indicators, and the core variables described earlier and their lags. As the main goal of this model is describing how inflation responds contemporaneous levels of macro variables and to their lags, one month and two-month lags of these variables were created. For every variable y_t, L1_x_t=$x_{t-1}$and L2_x_t=x_{t-2}.

Conditional quantile function is as below:

$$Q_{d\pi t}(\tau/x_{t-1})=x'_{t-1}\beta(\tau),$$

Where

$\tau\in(0,1)$ is the qunatile of interest, which in this case is 0.5 (median) and 0.9 (upper tail).

$Q_{d\pi t-}$ is τ-quantile of monthly inflation change

x_{t-1} is vector of regressors

β(τ) is τ specific quantile regression coefficient which will be estimated.

QR model estimates β(τ) by minimizing following assymetric check loss functions:

$\widehat{\beta_\tau}=arg\ min_b\sum_{t=1}^{T}\rho_\tau(y_t-x_t b)$, where

$$\rho_\tau(u)=u(\tau-I\{u<0\}).$$

This model was executed across different dataset- full dataset, pre 2016, post 2016 without inflation expectation and post 2016 with expectations and several β coefficients were estimated for each quantile τ.

As parameter vectors are estimated, the core empirical model of this thesis- the Quantile Vector Autoregression model (QVAR) was built. This model builds on quantile specific estimates given by QR and uses them to simulate inflation dynamics for several periods.

General QVAR formula is:

$$Q(d_{\pi t}/\mathcal{F}_{t-1},\tau)=\beta_{0,\tau}+\beta_{\tau}'X_{t-1}$$

$$Q_{d\pi t}(\tau/\mathcal{F}_{t-1})=\beta_{0,\tau}+\beta_{1,\tau}L1\_dlog\_fx\_w+\beta_{2,\tau}L2\_dlog\_fx\_w+\beta_{3,\tau}L1\_d\_policy\_rate+\beta_{4,\tau}$$

$$L2\_d\_policy\_rate+\beta_{5,\tau}L1\_ipi\_growth+\beta_{6,\tau}L2\_ipi\_growth+\beta_{7,\tau}L1\_dlog\_oil\_w+\beta_{8,\tau}$$

$$L2\_dlog\_oil\_w+u_{t,\tau}$$

This model gives an answer to question “What will inflation look like in the next 12 months if we are in a high-risk quantile- stress period, such as $\tau = 0.90$?”.

Quantile Impulse Response Functions (QIRFs)

After the QVAR model estimation, the next step is to analyze how inflation responds to different shocks specifically in the upper tail of its distribution, where risks are the highest.

In this paper QIRFs are calculated by simulation that follows the same logic as the QVAR model. To ensure that the model begins from real data, not just arbitrary values, the last two monthly inflation values are taken from the dataset and treated as the first lag L1_dπ and second lag L2_dπ. Then, all other variables including lags which were used in QVAR model are fixed at their most recent values. This supports the assumption that in the short term these variables can not respond immediately and hence can be treated as predetermined when simulating inflation risk.

The whichnlfation path under normal conditions whichs simulated using QVAR coefficients estimated on median quantile creatinf a sequence of 12 variables: $Q_{t+1}^{base}(\tau)$, $Q_{t+2}^{base}(\tau)$,..., $Q_{t+12}^{base}(\tau)$. The inflation path under shock scenarios in turn is simulated using the coefficients of the upper tail, when τ=90. For the first half of the year the lagged value of the shock variable is replaced by its empirical tail value, imposing stress scenario into the system. For example, L1_dlog_fx in FX depreciation shock is replaced by its value at the 90th percentile of FX depreciation. After each simulated month, the inflation lags are updated dynamically using the predicted values $L1\_d\pi,t+h= Q_{t+h}^{shock}(\tau)$ *and* $L2\_d\pi,t+h=L1\_d\pi,t+h-1$, while non-inflation variables remain fixed. This generates full 12 month quantile stress path.

Finally, Quantile impulse response at horizon *h* and quantile τ is defined as:

$$QIRF_{h(\tau)}=Q_{t+h}^{shock}(\tau)-Q_{t+h}^{base}(\tau),$$

where $Q_{t+h}^{shock}(\tau)$ is the predicted τ-quantile path of inflation when stress scenarios are applied and Q_{t+h}^{base} is the predicted τ-quantile path of inflation under normal conditions. This gives a step-by-step picture of how a shock to oil, the exchange rate, or the policy rate changes inflation risk. The fact that QIRFs are generated from the QVAR model results in autoregressive nature of inflation and the tail-driven behaviour of shocks being correctly captured.

Inflation at risk

Whereas the QIRFs provide dynamic and shock specific view of inlfation risk and show how external and domestic shocks propagate through the tail distribution of inflation, policymakers are also interested in getting a single measue that would be interpretable and can summarize tail outcomes and. Therefore, the next step was estimation of inflation-at-risk (IaR), which measures how high monthly inflation could get in the worst scenarios. The computation of IaR begins with

recalling the quantile regression model used in QVAR. For each tail level $\tau \in \{0.90;0.95;0.99\}$, the quantile model is estimated as:

$$d\pi_t = \beta_{0,\tau} + \beta_{1,\tau} L1\_dlog\_oil\_w + \beta_{2,\tau} L2\_dlog\_oil\_w + \ldots + \beta_{7,\tau} L1\_d\_policy\_rate + \beta_{8,\tau} L2\_d\_policy\_rate + \beta_{9,\tau} L1\_d\pi_t + \beta_{10,\tau} L2\_d\pi_t.$$

After the model estimation, predicted conditional quantiles are generated for every observation in the dataset:

$$\hat{Q}_\tau(d\pi_t | X_t) = X_t' \hat{\beta}_\tau.$$

These predicted values form an empirical distribution of "tail inflation" implied by the quantile model.

The Inflation-at-Risk is calculated as τ-quantile of this predicted distribution:

$$IaR_\tau = Quantile(\hat{Q}_\tau(d\pi_t | X_t), \tau).$$

This formula above gives a single threshold which represents the inflation rate expected only in adverse tail events with probability of $1-\tau$.

The IaR model is therefore:

$$Q_{0.90}(d\pi_t | X_{t-1}) = \beta_0 + \beta_1 L1\_dlog\_oil\_w + \beta_2 L2\_dlog\_oil\_w + \cdots + \beta_{10} L2\_d\pi\_yoy.$$

For the robustness check, these IaR thresholds are compared to historical extreme inflation outcomes. The same percentiles (90%, 95%, 99%) are computed from the actual realised monthly inflation changes.

$$Historical_\tau = Quantile(d\pi_t^{actual}, \tau).$$

This step simply checks whether the tail inflation numbers predicted by the model are similar to the inflation jumps that actually happened historically in Kazakhstan.

Stress Testing

After estimating Inflation-at-Risk (IaR), the next step is to examine how sensitive the inflation tail is to shocks that the key macroeconomic drivers could experience. This is done through two complementary procedures.

The first one is sensitivity stress test, the goal of which is analizing how inflation at risk changes when Kazakhstan's economy experiences large shocks. To conduct this test, firslty the baseline IaR model at the 90th conditional quantile of the inflation distribution is estimated. This model then serves as a benchamrk to measure changes under different shocks implied to lagged regressors of QVAR model that was defined earlier. These shocks are constructed as three-standard-deviation increases in the corresponding lagged variables (e.g., L1_dlog_fx_w+3σ, L2_dlog_oil_w+3σ), or as a 3 percentage-point persistence shock to lagged inflation (L1_d_π+3). For each stress scenario, the model is re-evaluated using the same 90th-quantile coefficients, and consequently new IaR measure is computed. So the difference between the stressed IaR and the baseline IaR now shows how sensitive tail inflation risk is to extreme movements caused by macroeconomic and financial shock.

The second one whichs historical stress test, whichh aims at evaluating how well the model would have predicted inflation risks during the actual crisises. Four economical crisis episodes were chosen for testing purposes: 2008 global financial crisis, 2015 oil-price collapse, 2020 COVID-19, and 2022 geopolitical shock. IaR model is re-estimated using only pre-crisis data for each episode and 90th quantile inflation outcomes for the crisis months are calculated. These

predicted inflation measures are then compared to actual inflation numbers during each crisis. This step is necessary to analyze whether the model is capable enough to anticipate high inflation risk before the crisis actually took place.

ECB-Style Multi-Month Quantile Stress Testing

Earlier, the QIRF section examined how a one-period shock moves the τ-quantile of inflation. In this part of the methodology, the European Central Bank (ECB) Global Financial Stability Review method is followed, which allows shocks to continue for several months, which in turn produces a multi-month stress direction of future inflation quantiles. The goal of this stress testing is to answer "how inflation behaves if Kazakhstan's economy remains in a crisis situation for several consecutive months?".

This part of the methodology directly builds on the earlier QVAR model. However, since Kazakhstan experienced major structural change at the end of 2015 which allowed free floating exchange rate regime and thus changed monetary transmission fundamentally making pre and post reform dynamics incompatible. That is why, this section used only post 2016 data for the analysis.

The stress test uses the same QVAR specification as before, where the τ-quantile of next month's inflation change, $Q(d\pi_{t+1}|F_t,\tau)$, is modeled as a function of lagged oil prices, exchange rates, industrial production, the policy rate, and inflation's own lags. Four stress scenarios are tested. The first one is setting lagged FX depreciation term $L1\_dlog(FX)_t$ is set to it's 90th percentile level. The second scenario is when $L1\_dlog(oil)_t$ is set to its 10th percentile, indicating crash value. The third is policy scenario, where $L1\_dpolicy\_rate_t$ to its tightening value-90th percientile. And the last one is joint-tail scenario which combines simulatenously all these three

stress scenarios. For each scenario, the model produces a 12 month forward behavior of inflation using $\tau = 0.90$ coefficients during the stress period and compares it against a baseline behavior generated with $\tau = 0.50$ coefficients. After, the stress impact is summarized using three metrics: the final-month effect which is the difference between the inflation prediction under shock scenarios and the baseline prediction at the last month of the simulation, the largest inflation deviation at any point during the year, and the cumulative effect across all twelve months. These measures provide a comprehensive overview of how multi month shocks shape the upper tail of inflation, allowing the stress-testing framework to capture risks that evolve and build over time.

Rolling Out-of-Sample Forecast Evaluation

The final part of the methodology compares QVAR with other models to evaluates how accurately the QVAR model can predict inflation tail. To do this, a rolling out of sample forecasting procedure was used.

The QVAR model is compared to 3 benchmarks: standard VAR, Random Walk which assumes inflation this month and next month is equal. Only post 2016 data was selected, where first 60 months are taken as initial estimation window. Then, the remaining months are split into two: training sample, which includes all observations from the start until t-1 and test observation which is the single month t.

Using the training sample, 4 different models were estimated. First one is QVAR, which is capable to produce direct prediction of 90th quantile inflation for the next month t. The second is standard VAR, which predicts the meand and the standard deviation of the next month's inflation. To convert this into a 90th quantile tail forecast, normal distribution is used:

$$Q_{0.90}^{VAR} = \mu_{t+1} + z_{0.90} \cdot \hat{\sigma}_{t+1},\ z_{0.90} = 1.2816$$

Third model is Random walk where 90th quantile is equal to the previous inflation value.

To accsees the forecast accuracy the pinball loss function was used. It is designed for quantile forecasting and penalizes underprediction of tail risks more heavily. Then, to evaluate the calibration, whether the predicted 90th quantile is exceeded approximately 10% of the time, the Kupiec coverage test is applied.

These methods of evaluation together provide a comprehensive assessment of how well the QVAR model works in predicting extreme inflation outcomes.

RESULTS

This results section presents the empirical results of the study completed so far and begins with the estimation of the baseline VAR model and its Cholesky-identified structural form, which together reveal how shocks propagate through Kazakhstan's macroeconomic system. Structural Shock Matrix was as follow:

Structural Shock Matrix (A):

	dlog_oil_w	**dlog_fx_w**	**d_policy_rate**	**ipi_growth**	**d_pi_yoy**
dlog_oil_w	1.0000	0.0000	0.0000	0.0000	0
dlog_fx_w	0.0786	1.0000	0.0000	0.0000	0
d_policy_rate	-0.0019	-0.0129	1.0000	0.0000	0
ipi_growth	0.0171	0.0489	0.3382	1.0000	0
d_pi_yoy	-0.0093	-0.0492	0.2647	0.0393	1

Table 1. Structural Shock Matrix (A)

This A-matrix shows that oil shock passes to the exchange rate immediately, with a positive coefficient equalling to 0.0786 indicating that higher global oil prices lead to appreciation of tenge. This is the expected result as Kazakhstan is an oil exporter. Policy rate in turn negatively responses to shocks in exchange rate, which simply shows the tendency of National Bank tightnening monetary policy when exchange rate depreciates. The policy rate shock positively transmits to IPI and to inflation simultaneously, which reflects Kazakhstan's exchange-rate and credit-channel sensitivities. Inflation experiences higher response to shocks in exchange rate, while responses to shocks in oil price and real activity is comparatively weaker. Overall, the structure is consistent with theoretical and empirical predictions and verifies the Cholesky ordering employed in the benchmark VAR.

After the matrix, the IRF graphs were generated. The most important three of them are discussed.

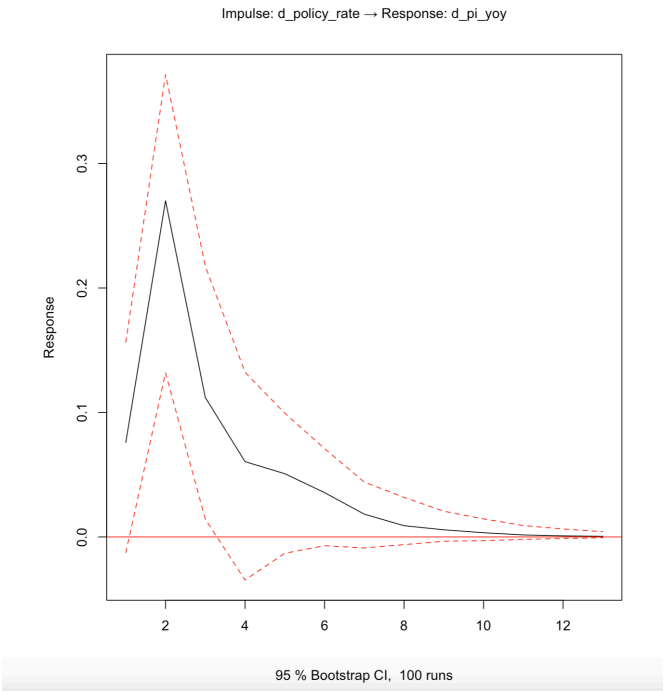

Figure 1. Impulse Response of Inflation to a Monetary Policy Shock

First graph shows the dynamic effect of a monetary policy shock on inflation over the next 12 months. X-axis here horizon in months after the shocks, which is between 1 and 12, whereas Y-axis indicates change in monthly year-over-year inflation. Black solid line is estimated IRF- how inflation responds to 1 standard deviation increase in the policy rate. Red dashed lines are 95% boostrap confidence intervals. They show the range of possible outcomes the model considers realistic. The pattern of the graph shows that the monetary tightening increases the inflation sharply during first 2 months. This is considered to be normal for emerging markets like Kazakhstan because structural characteristics like partial dollarization, cost-push inflation, and backward-looking price formation cause the immediate effects of interest-rate increases to raise prices before the disinflationary effects appear. After this initial spike, inflation gradually decreases to zero, which indicates that policy tightening has no permanent effect on inflation.

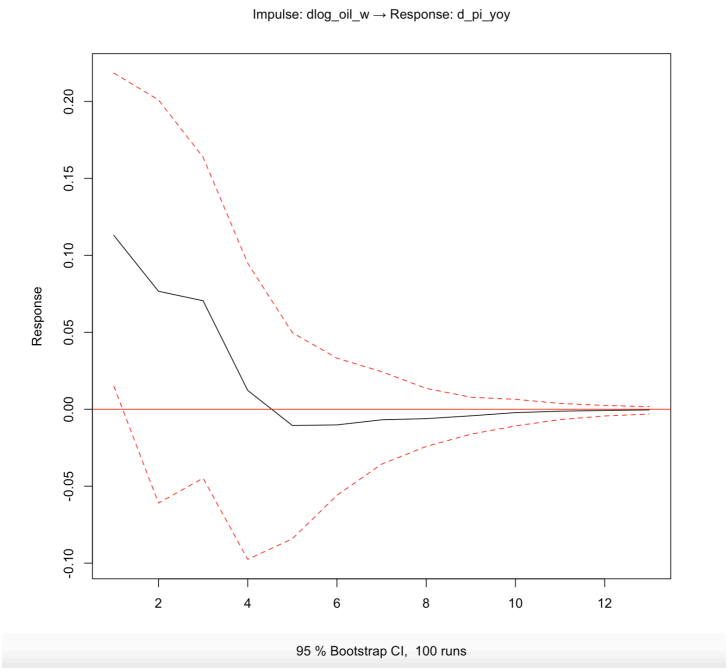

Figure 2. Impulse Response of Inflation to an Oil Price Shock.

This graph shows the inflation behavior over 12 months horizon after one standard deviation shock was imposed to global oil prices. In the first moths inflation reacts immediately and positively, even though the increase is small. The response is also temporary, as inlflation returns to its baseline level within 4-6 months. As confidence intervals are wide and zero for most horizons, it can be interpreted that the effect is not statistically strong. This pattern is typical for Kazakhstan, as oil prices influence inflation not directly but trough income, government spending, exchange rate etc.

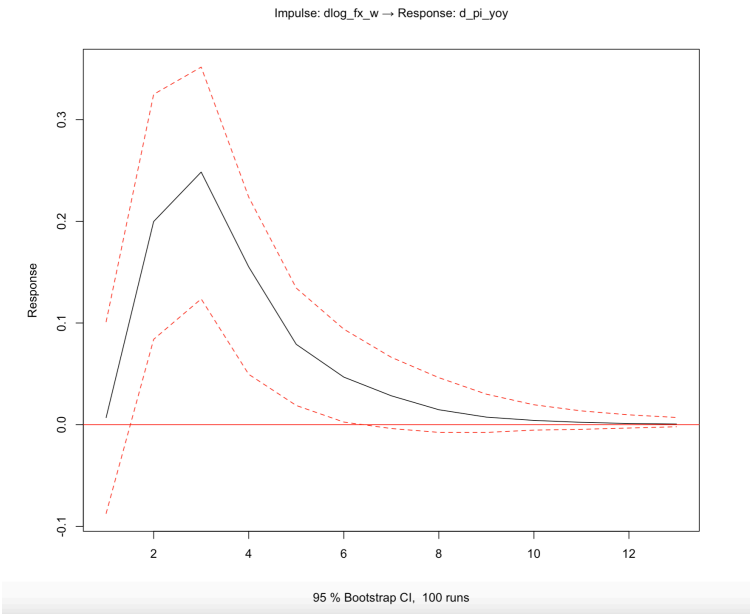

Figure 3. Impulse Response of Inflation to an Exchange Rate Depreciation Shock.

The last figure represents how inflation reacts over 12 months to a positive one standard deviation exchange rate depreciation shock. Depreciation of tenge lead to immediate increase in inflation, peaking around 2-3 month after the shock was imposed. The magnitude of the effect is larger compared to response to oil price shocks or policy tightening. This indicates that the exchange rate is the key driver of short-term inflation in Kazakhstan. Even though after reaching its peak, the inflation response gradually declines, it still remains positive for several months,

which reflects the typical lagged adjustment of import prices, production costs, and consumer prices. By around 8–10 months, the effect begins to fade and converges back to zero, showing the strong but short-term nature of the impact.

IRF results fully coincide with Cholesky ordering used in SVAR model. Oil shocks, which are ranked first in the ordering, have only weak and short-lived impact on inflation, whereas exchange-rate shocks, which are ranked second, have the maximum pass-through to prices, exactly as it was suggested by the recursive structure. Monetary policy shocks, which come after the exchange rate in the ordering, also affect inflation but only briefly and with a much smaller impact. This pattern fits well with how Kazakhstan's economy actually behaves and supports the idea that the Cholesky ordering reflects the true short-run relationships among the variables.

After establishment of the baseline dynamic relationships among macroeconomic variables through the VAR and SVAR impulse responses, this part moves to the baseline quantile regression models, which firstly shows how inflation responds under typical conditions ($\tau = 0.5$). These median estimates provide the reference point needed before turning to the more asymmetric and tail-sensitive quantile results. For theis purpose, 3 different tails were generated.

```
tau: [1] 0.5

Coefficients:
                               coefficients lower bd upper bd
(Intercept)                     -0.00877     -0.07782  0.06659
L1_dlog_fx_w                     0.03149      0.02907  0.05164
post_reform_dummy                0.01066     -0.07813  0.14763
L1_d_policy_rate                 0.36296      0.12538  0.68810
L2_d_policy_rate                 0.28336      0.08100  0.37942
L1_ipi_growth                   -0.01957     -0.03728  0.00381
L2_ipi_growth                    0.00709     -0.01499  0.01552
L1_dlog_oil_w                    0.00460     -0.00470  0.01011
L2_dlog_oil_w                   -0.00018     -0.00179  0.01110
L1_dlog_fx_w:post_reform_dummy -0.02526     -0.08483  0.01505
```

Table 2. Median Quantile Regression with FX Pass-Through Interaction Term (Full Sample, 2006–2024)

The first table represents full sample interaction table, which estimates inflation at the median over 2006-2024 and includes the interaction term L1_dlog_fx_w × post_reform_dummy. This is necessary to test whether FX pass-through experienced any changes after the regime shift.

The coefficient of the first lag depreciation term equallin to 0.03149 confirms the meaningful FX pass through at the median, which means that changes in exchange rate gets transmitted into domestic inflation in a statistically significant and economically relevant way. According to numbers in the table, policy rate has the largest coefficients, however, this only indicates the central bank's tendency to adjust rates in response to emerging inflation pressures, rather than claiming that policy tightening directly generates inflation. In contrast, real activity and oil prices have coefficients around zero, indicating no short-term impact to inflation on average. The interaction term capturing post 2016 is negative, suggesting weakening of FX pass-through after the regime shift, although the effect was found to be not statistically significant.

Given that the post-2016 period corresponds to a structurally different monetary environment, further analysis focuses on two quantile regression specifications estimated solely on post-reform data. The first regression is without inclusion of inflation expectations, whereas the second one includes this indicator.

```
tau: [1] 0.5

Coefficients:
                  coefficients lower bd upper bd
(Intercept)        0.02379     -0.02905  0.11102
L1_dlog_fx_w       0.01800     -0.04484  0.04746
L2_dlog_fx_w      -0.01177     -0.05211  0.04165
L1_d_policy_rate   0.43590      0.08669  0.59325
L2_d_policy_rate   0.42730      0.03501  0.49967
L1_ipi_growth     -0.00815     -0.02942  0.02289
L2_ipi_growth      0.00459     -0.01174  0.03030
L1_dlog_oil_w      0.00463     -0.01453  0.01175
L2_dlog_oil_w     -0.00134     -0.00482  0.00648
```

```
tau: [1] 0.5

Coefficients:
                  coefficients lower bd upper bd
(Intercept)        0.68134     -0.08577  1.08601
L1_dlog_fx_w      -0.02522     -0.06061  0.03886
L2_dlog_fx_w      -0.03768     -0.08139  0.01395
L1_d_policy_rate   0.37102      0.06199  0.52053
L2_d_policy_rate   0.41145      0.03851  0.52577
L1_ipi_growth     -0.01642     -0.04132  0.01454
L2_ipi_growth      0.00516     -0.04394  0.01921
L1_dlog_oil_w     -0.00425     -0.01914  0.00665
L2_dlog_oil_w     -0.00339     -0.00800  0.00325
L1_inflation_exp   0.01537     -0.07076  0.07467
L2_inflation_exp  -0.07160     -0.13754  0.01644
```

Tables 3-4. Post-2016 Median Quantile Regression Results (τ = 0.50), With and Without Inflation Expectations

The post-2016 median quantile regression, reported in the left Table, show the FX coefficients being small (0.018 anf -0.1177), with confidence intervals fluctuiating around 0. This emphasizes that after the regime shift, under floating regime, exchange-rate movements no longer exhibita statistically significant impact on typical monthly inflation. Other variables behave similarly to the results obtained using full sample: domestic demand and oil price effects remain minor, while the policy rate terms continue to show the strongest and most consistent association with median inflation.

As it can be noticed from the second table, including survey-based inflation expectations does not materially change the underlying drivers of median inflation. Hence, it can be concluded that once macroeconomic variables are included, inflation expectations does not add additional information for explaining median inflation behavior.

Next, in order to further evaluate whether these median effects remain stable across the full distribution of inflation outcomes, the analysis turns to the quantile regression coefficient profiles represented in the form of plots.

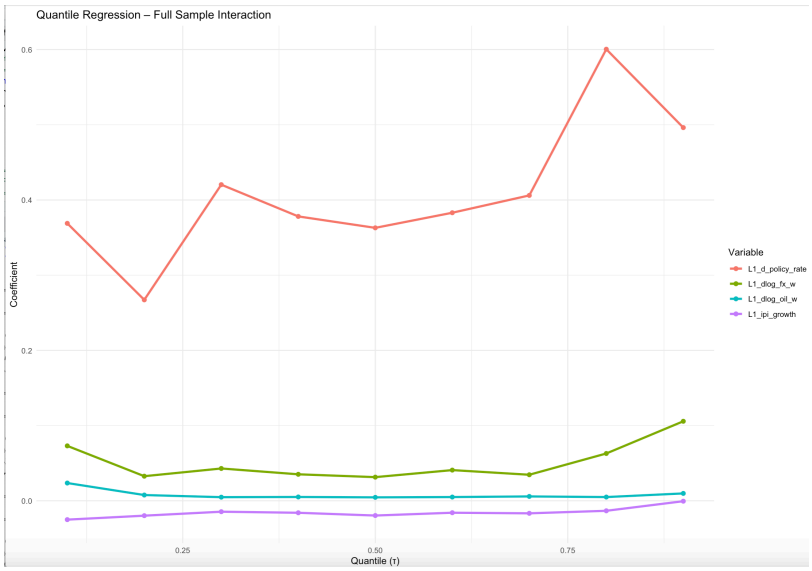

Figure 4. Quantile Regression- Full sample interaction.

The first graph describes how the impact of key macroeconomic drivers on monthly inflation changes across different points of the inflation distribution ($\tau = 0.1$ to 0.9), not just at the average. The red line shows the behavior of the policy rate (L1_d_policy_rate). Its coefficient is positive across all quantiles, peaking at the upper tail (rising from about 0.35 at $\tau = 0.1$ to nearly 0.60 at $\tau = 0.9$). Such pattern suggests that monetary policy reacts most aggressively during high-inflation periods, which is intuitive knowing Kazakhstan's counter inflationary policy stance. In contrast, the coefficients of exchange rate, oil prices and real activity fluctuates around zero across all quantiles. These results suggest that shocks in exchange rate, external commodity and domestic demand does not shape short term inflation behavior, once policy reactions are controlled for.

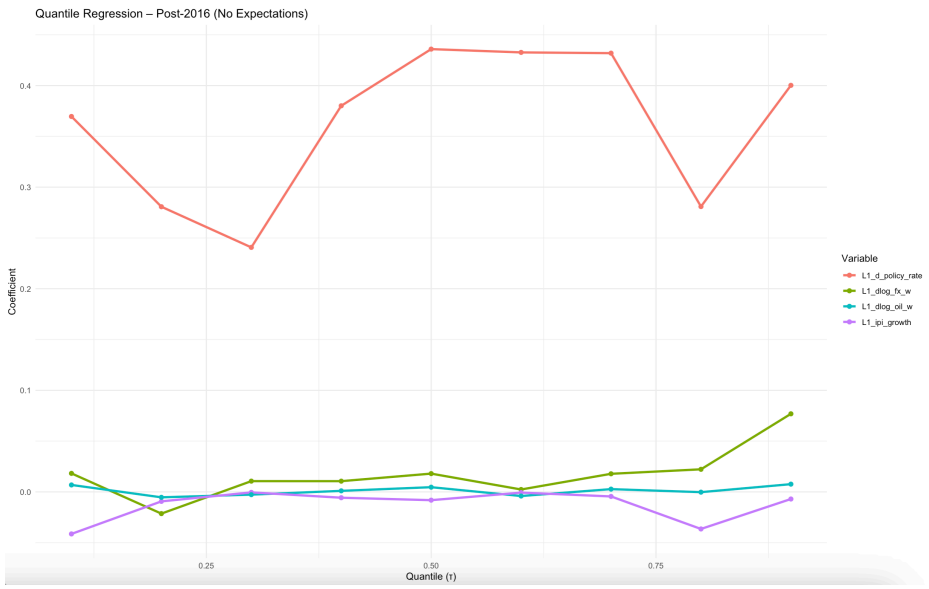

Figure 5. Quantile regression- Post-2016 (No expectations)

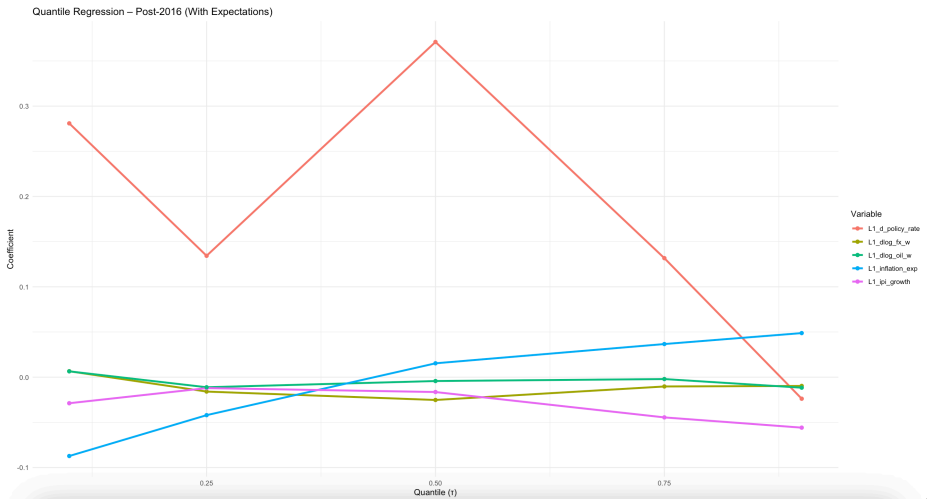

Figure 6. Qunatile regression-Post 2016 (With expectations)

Both graphs above show the pattern after the regime shift. The policy rate still remains dominant, with coefficients being much larger than other variables. The main difference between 2 graphs is smoothness and magnitude of the coefficient of this driver. When expectations are included, the policy rate becomes slightly smaller and smoother. This aligns with the notion that some of the change in inflation is already predetermined by forward-looking beliefs. However, the

general shape of the policy-rate curve is stll similar in both models and the fact that expectations themselves remain close to zero across all quantiles confirms that including expectations does not materially change the core transmission mechanism. The remaining variables cluster around zero in both graphs, same as in full sample.

	tau	R1_noexp	R1_exp	R1_gain
1	0.10	0.08088826	0.1697929	0.088904669
2	0.25	0.06695394	0.1524900	0.085536042
3	0.50	0.09319692	0.1013261	0.008129208
4	0.75	0.13432498	0.1258484	-0.008476631
5	0.90	0.18608959	0.1826196	-0.003469995

Table 5. Pseudo-R^2 Comparison Across Quantiles With and Without Inflation Expectations.

For the further asessment of inflation expectations to model performance, the pseudo-R^2 comparison across quantiles were included. The results show that at the lower distribution of inflation, addition of inflation expectations to the post 2016 model improves model fit by approximately 0.09. At the median and higher tails, the effect is minimal. Therefore, it can be completed that inclusion of this variable helps to explain the inflation only when inflation is at its lower distributions.

The quantile-specific coefficient plots described static relationship by showing the inflation behavior with shocks at different points of the distribution. The work now shifts to the investigation of dynamic dimension by investigating how shocks transmit at the upper tail of the inflation distribution using quantile impulse-response functions. On this step 2 samples were used: full sample and post 2016 sample without inclusion of inflation expectations. The reason behind excluding inflation expectations is that the post-2016 QIRFs with inflation expectations could not be computed as expectations sample was too small to support the full lag structure of

the quantile VAR. Moreover, the quantile regression results already showed that inflation expectations are statistically insignificant across high quantiles, making usage of sample with expectations unnecessary. Important to add that IPI was excluded as shocks variable in this part, as it is more about endogenous activity indicator rather than a structural or exogenous driver.

	Horizon	Oil_Shock	Oil_SE	FX_Shock	FX_SE	Policy_Shock	Policy_SE
1	1	0.0037	0.0116	0.0535	0.0823	0.2676	0.2016
2	2	0.0155	0.0160	0.0922	0.0592	0.1670	0.2167
3	3	-0.0027	0.0166	0.0105	0.0316	0.2313	0.1742
4	4	0.0014	0.0094	-0.0027	0.0300	0.0707	0.1447
5	5	0.0041	0.0090	0.0174	0.0285	0.0946	0.2220
6	6	0.0129	0.0083	0.0084	0.0211	0.1544	0.1480
7	7	0.0126	0.0088	-0.0069	0.0224	0.1379	0.1544
8	8	0.0021	0.0123	-0.0197	0.0199	0.2911	0.1375
9	9	-0.0003	0.0196	-0.0048	0.0233	0.2171	0.1713
10	10	-0.0021	0.0128	0.0240	0.0373	-0.1025	0.2032
11	11	-0.0161	0.0172	-0.0348	0.0554	0.1492	0.2080
12	12	-0.0073	0.0157	-0.0478	0.0640	0.0710	0.3300

Table 6. Full sample QIRF Table τ = 0.90.

	Horizon	Oil_Shock	Oil_SE	FX_Shock	FX_SE	Policy_Shock	Policy_SE
1	1	-0.0058	0.0144	0.0166	0.0390	0.2557	0.2454
2	2	0.0176	0.0171	0.1020	0.0519	0.3327	0.2579
3	3	0.0023	0.0200	0.0301	0.0448	0.4361	0.2754
4	4	0.0059	0.0135	0.0032	0.0399	0.3991	0.2534
5	5	0.0033	0.0119	-0.0042	0.0360	0.2961	0.2248
6	6	0.0068	0.0102	0.0020	0.0465	0.3829	0.1598
7	7	0.0135	0.0105	-0.0256	0.0496	0.4731	0.2677
8	8	0.0050	0.0146	-0.0201	0.0534	0.3624	0.2938
9	9	0.0045	0.0174	0.0025	0.0582	0.3104	0.2798
10	10	0.0177	0.0184	0.0492	0.0956	0.2213	0.3470
11	11	0.0075	0.0140	-0.0159	0.0867	0.0431	0.3187
12	12	-0.0003	0.0121	-0.0357	0.0649	-0.0417	0.2933

Table 7. Post-2016 Quantile Impulse Response Functions (QIRFs) at τ = 0.90, Without Inflation Expectations

The full-sample and post 2016 QIRFs at τ = 0.90 show the behavior of extreme inflation outcomes over 12 months horizon when different macroeconomic shocks are applied. In both

cases, policy rate coefficients are the largest compared to other variables. They found to be between 0.07 and 0.29 in full sample and between 0.3 and 0.47 in post 2016 across quantile. These findings are consistent with what was revealed by quantile regression model earlier- high inflation leads to contemporaneous adjustments in the policy rates, reinforcing interpretation claiming that policy rates act primarily as a reactive instrument, rather than causal driver of inflation. FX shocks in turn generate modest and unstable responses- positive in the short run but turn negative at several horizons, which confirms the previous finding that exchange-rate pass-through is weaker in periods of significant inflation. Oil price shocks remained small and economically insignificant in both samples.

To better understand Kazakhstan's inflation tail risk, the distribution of predicted conditional quantiles was created. The model suggests that monthly inflation changes at the mean are relatively moderate -the mean predicted change is 0.57 p.p. with the median 0.50 p.p. However, the upper tail in contrast becomes thicker with 90th percentile reaching 1.10 p.p., the 95th percentile increasing to 1.48 p.p., and the 99th percentile reaching 3.08 p.p. This reveals Kazakhstan's highly skewed indicating that Kazakhstan's inflation process is highly skewed with large movements occurring in the tail of the distribution.

The next step estimates the Inflation-at-Risk, the extent of monthly inflation spikes under adverse macroeconomic conditions. According to the model estimations:

- at the 90^{th} percentile, inflation could rise by 1.098,

- at the 95^{th} percentile the risk rises to 1.85

-at 99^{th} percentile immediate spikes to over 6.4 percentile points in a single month.

These obtained results were further compared to historical extremes. According to the past:

-the 90th percentile of actual monthly inflation was 0.7 p.p

- the 95th percentile was 1.07 p.p

- 99th percetile was equal to 3.08 p.p.

The thresholds predicted by QVAR model are slightly higher compared to historical, but this was expected as QVAR tries to capture conditional tail risk-inflation that could occur. These findings confirm that Kazakhstan's inflation tail risk is large and assymetric and at the same meanigfully predictable using quantile-based methods.

Following the estimation of Inflation-at-Risk, its response to artificial macroeconomic shocks is investigated. The IaR at 90th percentile equaling 1.098 p.p served as baseline in this part and 3 large shocks were applied.

```
Testing shocks on lagged regressors (τ = 0.90 IaR):
  Baseline IaR(90%) = 1.098
  FX_3sd_shock      : Shocked_IaR=2.049  Change=0.951
  OIL_3sd_shock     : Shocked_IaR=1.621  Change=0.523
  INFL_PERSIST_3pp  : Shocked_IaR=2.737  Change=1.639
```

Table 8. Inflation-at-Risk Sensitivity to Macro-Financial Shocks (τ = 0.90)

The first shock was 3 standard deviation increase in exchange rate. This shock increased the Inflation at Risk by 0.951, showing that large depreciation shock increases the probability of extreme inflation outcome. Identical shock was applied to oil prices, and caused more modest rise of 0.523 percentage points, which is consistent with the generally weak oil-inflation linkage observed throughout the study. Inflation persistent was found to be the most powerful driver of the inflation, as adding 3 percentage points to lagged inflation pushed IaR by 1.639 pp. In general, results suggest that extreme inflation outcomes in Kazakhstan are most sensitive to

internal inflation dynamics and even though exchange rate was found to be insignificant as macroeconomic driver, large shocks applied to this variable still increase the tail risk of inflation.

Artficial shock testing was then followed by historical stress testing. The goal of this test was to evaluate whether IaR model trained only on pre-crisis data, can anticipate the scale of inflation risk observed during real world shocks. For testing purposes, 4 historical crisis periods were selected: 2008 Global Financial Crisis, 2015 oil price crash, 2020 COVID-19 period and 2022 Russian Invasion period.

```
Historical Crisis Performance (τ = 0.90 IaR, model estimated on pre-crisis data):
  2008_GFC            : Actual=0.080  Predicted=4.411  Error=4.331
  2015_Oil_Crash      : Actual=1.580  Predicted=1.685  Error=0.105
  2020_COVID          : Actual=0.370  Predicted=2.432  Error=2.062
  2022_Russia_Invasion: Actual=1.600  Predicted=1.931  Error=0.331
```

Table 9. Historical Crisis Backtesting of Inflation-at-Risk (τ = 0.90).

According to the numbers predicted, the predictive power of this model was strong during 2015 oil crash crisis and 2022 Russian Invasion. In both episodes, predicted high-inflation outcomes closely matched what actually occurred, having errors around zero (0.105 and 0.331). However, the model overestimated the inflation risk during Global financial crisis and 2020 Covid-19. Results are totally expected and interpretable. 2015 and 2022 episodes are externally driven commodity or exchange rate shocks that translates into inflation tail risk. What comes to crises of 2008 and 2020, during these periods there were severe policy interventions from National Bank. In 2008 Kazakhstan still had fixed exchange rate regime, which absorbed much of the external shock and thusprevented immediate pass-through to domestic prices. That is why inflation did not spike up in a way the model predicted. The difference between actual and predicted Inflation-at-Risk in 2020 can also be partly explained by the extraordinary policy interventions undertaken

during the COVID-19 crisis. The National Bank of Kazakhstan conducted large-scale FX interventions (over USD 3 billion in March alone) to stabilize the tenge. In addition, NBK coordinated FX conversion operations with major quasi-sovereign entities, including the Unified Accumulated Pension Fund (UAPF) and firms within the Samruk-Kazyna group, to increase FX liquidity in the domestic market. NBK also implemented an emergency 2.75 percentage-point policy rate hike to curb depreciation-driven inflation pressures. At the same time, the government introduced extensive administrative price controls and anti-crisis fiscal subsidies, which further altered the inflation path in ways that standard econometric models may not fully capture. These actions alltogether artificially suppressed the usual exchange-rate and supply-shock transmission mechanisms captured by the QVAR.

After evaluating the model's performance during real-world crises, this part turns to a more forward-looking stress-testing framework. The hostorical backtesting method demonstrated how the IaR responded to previous shocks, whereas the ECB-style multi-month simulations were intended to show how would inflation behave if Kazakhstan were hit by new extreme shocks today. Each "shock" in this case corresponds to forcing one or more lagged macroeconomic variables in the QVAR model to their extreme quantile values for 6 consecutive months.

```
Scenario: FX_tail
  Last-month effect: 1.993 pp
  Max deviation:     1.995 pp
  Cumulative effect: 20.603 pp

Scenario: OIL_tail
  Last-month effect: 1.993 pp
  Max deviation:     1.995 pp
  Cumulative effect: 21.495 pp

Scenario: POLICY_tail
  Last-month effect: 1.994 pp
  Max deviation:     1.995 pp
  Cumulative effect: 20.422 pp

Scenario: JOINT_tail
  Last-month effect: 1.993 pp
  Max deviation:     1.995 pp
  Cumulative effect: 20.839 pp
```

Table 10. Multi-Month Quantile Stress-Testing Results for FX, Oil, Policy, and Joint Tail Scenarios ($\tau = 0.90$).

Across all scenarios the model produced almost similar outcomes: 1) by the end of the simulation, inflation in the upper tail increased by approximately 2 p.p. above baseline, with maximum deviation of roughly 2 p.p. at any point. The cumulative pressure over the 12-month horizon is also large and broadly comparable across shocks- around 20 to 21.5 percentage points.

This result is a significant structural point rather than a model failure: inflation persistence dominates tail outcomes, and the τ= 0.90 QVAR offers minor coefficients to all the macro shocks. Therefore, once inflation reaches an unfavorable high-inflation regime, the tail distribution is no longer sensitive to the initial shock's character because future inflation values are almost entirely dependent on past values. This outcome serves as additional confirmation the earlier findings that Kazakhstan's inflation tail risk is primarily persistent rather than shock-driven.

As final step in this thesis rolling out-of-sample forecasting evaluation was done.

Model	**Pinball Loss**	**Hit Ratio**	**Kupiec p-value**
QVAR	**0.18911**	**0.851**	**0.2933**
VAR(2)	1.78311	0.064	0.0000
Random Walk	0.27170	0.532	0.0000

Outcome of this out-of-sample evaluation shows that QVAR model is best model that is able to forecast upper tail inflation forecasts. Pinball loss of QVAR is substantially lower compared to other methods, which means that its tail specific forecasts are more accurate. The Kupiec coberage test also confirms the reliability of this model as its hit ratio of 0.851 is close to the desired 90% level, and its p-value of 0.2933 indicates no statistical rejection of correct coverage. By contrast, classical VAR models perform extremely poorly, experiencing very high loss and a severe under-coverage problems. The random walk also fails to provide adequate tail-risk forecasts. As conclusion of this methodological part, these results strengthen the idea that quantile-based methods capture the dynamics of inflation tail risk far more effectively than traditional linear forecasting approaches.

CONCLUSION

This thesis shows that state-dependent dynamics in Kazakhstan inflation: on the one hand, it is typical of the traditional VAR model that the exchange-rate depreciation should promote average inflation, but the QVAR model produces the opposite result in the upper tail. Inflation in the 90

th percentile is mainly self-perpetuated other than externally shock due. This difference explains why inflation in the 2015 currency crisis and the 2022 commodity surge proved to be so persistent than the mean-based models predicted—policy makers whose goals are to stabilize the exchange rate had to deal with the issue of inflation momentum that kept growing even when the initial shocks had calmed.

The practical usefulness of this framework is evident in the form of the particular outcomes. The six-month-long multi-month simulations under ECB style show that a six-month period of continuous exchange-rate depreciation of the exchange rate moves upper-tail inflation roughly 2 percentage points above base in a year, adding up to more than 20 percentage points of excess inflation pressure. This quantification turns the deliberations of the monetary policy committee radically different. Instead of posing the question of whether inflation will be at its historical average, the policy makers can now pose the question of whether there is a likelihood of us witnessing double digit inflation and causing a loss of credibility.

This difference was a critical one in 2015. The ability of the model has been tested historically: using only pre-crisis data to train it, QVAR was able to forecast 2015 crisis inflation with an error of 0.105 percentage points and 2022 crisis inflation with an error of 0.331 percentage points. Both episodes were being systematically underestimated by the traditional VAR models. If NBK had QVAR-based risk assessments in August-September 2015 with big probability of severe inflation spike that eventually hit 14 percent, the monetary policy committee would have justified the use of stronger preemptive tightening that would have avoided the de-anchoring of inflation expectations that extended the crisis into 2016.

Such aggressive action is justified as shown by the distributional analysis. On the one hand, median monthly inflation change is not high; it is moderate with 0.50 pp, and the 95 th percentile

value is 1.85 pp or almost four times greater. Spikes on the 99 th percentile to 6.4 pp in one month. Conventional forecasting completely ignores this asymmetry and monetary policymakers are only aware of tail risks that dictate whether inflation remains anchored or takes off target.

The results allow making three definite improvements to the inflation-targeting regime at NBK.

To begin with, inflation-at-risk indicators offer early warning signals which are actionable. When the forecasted 95th percentile is significantly below the median predictions (e.g. 1.85 pp as the median is at 0.50 pp), then the economy is transitioning to the high-risk regime and persistence mechanisms are turning on. The monetary policy committee, at this stage, ought to contemplate emergency policy responses in order to avoid the full entry of inflation into the persistence-dominated stage whereby transmission of policy becomes considerably weak.

Second, the stress-testing module allows strong scenario planning of the monetary policy. The committee is able to consider the entire distribution of possible outcomes given the various shock sequences, rather than having to just look at single-point forecasts or make arbitrary assumptions. The realization that persistent volatility of the commodity prices coupled with the exchange-rate pressure forms the tail risks that compound each other, up to more than 20 percentage points of excess pressure in one year, can be used to gauge the right amount of policy response and time-frame.

Third, tail-risk prediction is clearly superior under forecasting assessment in the credibility of inflation-targeting. QVAR framework has a 90 percent coverage accuracy and a much lesser pinball loss compared to VAR and random walk models. Mean-based forecasts have been found to be useful in normal times but they are systematically understated in times of high uncertainty,

precisely when central bank credibility is highly valued and de-anchoring of inflationary expectations that proved very expensive in the years following the crisis of 2015.

This thesis offers a more realistic framework compared to the average results by switching to full risk distributions, which are more realistic in the realities of small open economies that suffer the vagaries of the external environment. In the case of Kazakhstan, where an exchange-rate movement, a commodity cycle, and domestic persistence are often at play, it is critical to know the upper tail that can be used to achieve success in inflation targeting and the sustainability of the monetary policy framework.

REFERENCE

Adrian, T., Boyarchenko, N., & Giannone, D. (2019). *Vulnerable growth*. American Economic Review, 109(4), 1263–1289.

Ahmed, R., Khan, A., & Rehman, M. (2024). *Inflation targeting and inflation tail risks in emerging economies*. Journal of Monetary Economics, 139, 45–62.

Banerjee, R., Mehrotra, A., & Zampolli, F. (2024). *Inflation-at-Risk: Drivers and dynamics across advanced and emerging markets*. BIS Working Paper Series.

Banbura, M., Hubrich, K., & Lenza, M. (2024). *Quantile Structural VARs for policy stress testing*. Journal of Econometrics.

Chavleishvili, M., & Manganelli, S. (2020). *Forecasting and stress testing with quantile vector autoregressions*. ECB Working Paper No. 2330.

Huang, Y., & Zhang, J. (2023). *Inflation asymmetries and global commodity shocks*. Journal of Applied Econometrics, 38(5), 743–769.

Kilian, L., & Vigfusson, R. (2017). *The role of oil price shocks in causing recessions*. Review of Economics and Statistics, 99(5), 1–15.

Korobilis, D., Mumtaz, H., & Theophilopoulou, A. (2021). *Inflation risks and monetary policy regimes: A quantile regression approach*. Journal of Economic Dynamics & Control, 130, 104197.

Makabe, I., & Norimasa, K. (2022). *A term structure of inflation-at-risk*. Journal of Macroeconomics, 71, 103426.

Montagnoli, A., & Napolitano, O. (2022). *Oil shocks, inflation asymmetries, and quantile VAR evidence*. Energy Economics, 110, 106004.

Queyranne, M., Lafarguette, R., & Johnson, K. (2022). *Inflation-at-Risk in the Middle East and Central Asia*. IMF Working Paper WP/22/168.

Schule, A. (2020). *Uncertainty shocks and the business cycle: A quantile structural VAR approach*. Journal of Macroeconomics, 64, 103215.