Modeling Latent Volatility Regimes in Daily WTI Crude Oil Returns: A POMP Analysis with ARIMA and GARCH Benchmarks

1 Introduction

Crude oil is a central commodity in the global economy, and its price dynamics respond quickly to changes in supply conditions, transportation frictions, and broader macroeconomic uncertainty. These features make oil returns a natural object of study when the goal is to understand changing volatility across more stable and more turbulent market conditions.

In this project, we study daily returns on West Texas Intermediate (WTI) crude oil and ask whether their dynamics can be described by a small number of latent volatility states. Our focus is on three features of the series: heavy tails, volatility clustering, and abrupt changes in market conditions. These features motivate going beyond a single linear Gaussian specification.

Our project also relates to earlier course work on crude-oil time-series modeling, but shifts the focus to daily WTI returns and latent volatility regimes (STATS 531 Final Project Group 2022). Our analysis proceeds in three steps. We first describe the data and examine preliminary diagnostics. We then fit ARIMA and AR(1)-GARCH(1,1)-t benchmark models. Finally, we estimate a partially observed Markov process (POMP) model. The POMP model improves substantially on the ARIMA baseline, while GARCH remains the strongest benchmark in terms of AIC. We therefore interpret the main contribution of the POMP model as its latent-state structure and economic interpretability rather than uniformly superior fit across all baselines.

2 Data

Our analysis uses daily spot prices for West Texas Intermediate (WTI) crude oil from the FRED series DCOILWTICO (Federal Reserve Bank of St. Louis 2026).

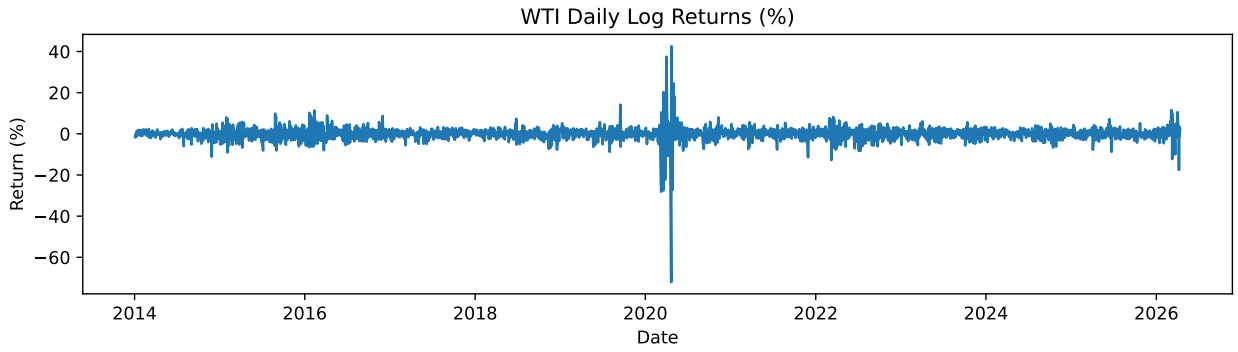
$$r_t = 100 \times \Delta \log(P_t),$$

and use this return series throughout the paper.

This transformation shifts the analysis from long-run level movements to short-run fluctuations, which is more appropriate for our focus on heavy tails, volatility clustering, and latent changes in market conditions. Using the same return series throughout also ensures that ARIMA, GARCH, and POMP are compared on a common empirical basis.

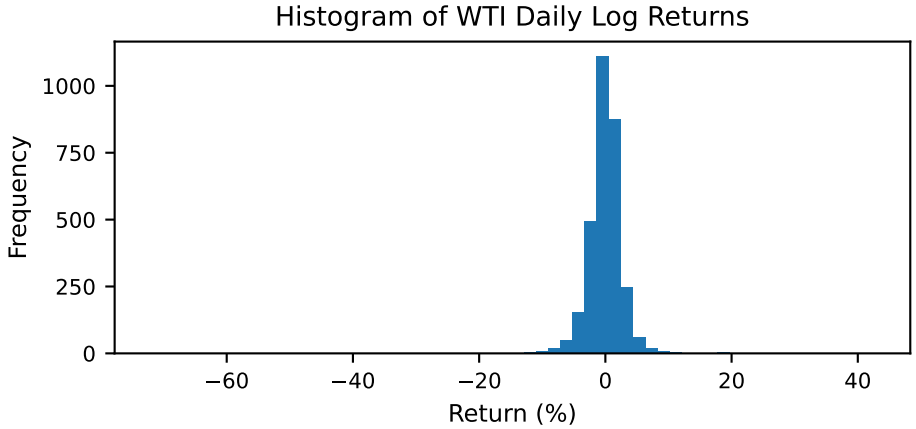
3 Preliminary Analysis

We begin with a set of exploratory diagnostics designed to clarify which features of the WTI return series require explicit modeling. In particular, we examine the price series, the return series, the marginal distribution of returns, and autocorrelation patterns in both returns and squared returns. These diagnostics are useful because they indicate whether the main structure of the data lies in the conditional mean, the conditional variance, or both.

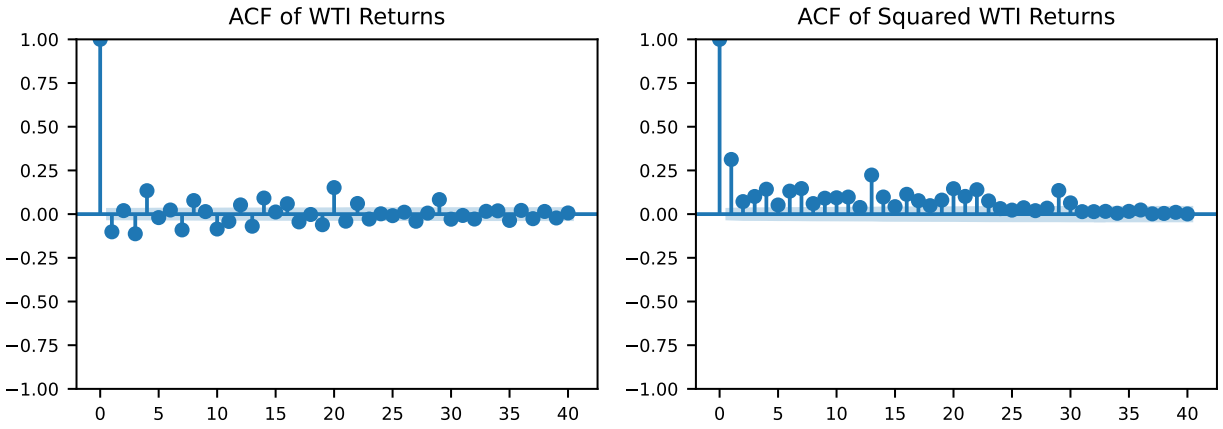


The price plot provides a broad overview of the sample and shows that crude-oil prices move through visibly different market environments over time. However, for the purposes of statistical modeling, the return plot is more informative. In the return series, periods of relatively modest day-to-day variation alternate with bursts of much larger movements, suggesting that volatility is not constant over time. This feature already points toward conditional heteroskedasticity as a central characteristic of the data.

We next examine the empirical distribution of returns using a histogram. Relative to a simple Gaussian benchmark, the return distribution appears to be heavy-tailed, reflecting the presence of rare but unusually large daily shocks. This is an important consideration for model specification, since a normal error distribution may understate the probability of extreme events in the oil market.



Finally, we consider autocorrelation diagnostics. The autocorrelation in raw returns is comparatively weak, which is typical for daily financial return series and suggests that linear mean dynamics alone are unlikely to explain much of the dependence in the data. By contrast, the autocorrelation in squared returns is more persistent, indicating that dependence is expressed more clearly through the volatility process than through the conditional mean. Taken together, these preliminary findings motivate the baseline models used in the next section: ARIMA as a benchmark for linear dependence in returns, and GARCH as a benchmark for time-varying volatility and heavy-tailed innovations. They also provide the empirical justification for the richer POMP specification developed later in the paper.



4 Baseline Time-Series Models

Before introducing the POMP model, we fit two standard benchmark specifications. The first is an ARIMA model, which serves as a conventional linear Gaussian baseline for the dynamics of the return series (STATS 531 2026c). The second is an AR(1)-GARCH(1,1)-t model with Student-t innovations, which provides a more appropriate benchmark when the main dependence in the data is expressed through time-varying volatility rather than the conditional mean.

The role of the ARIMA model is primarily diagnostic. Because the preliminary analysis suggests that autocorrelation in raw returns is comparatively weak, ARIMA is not expected to provide a fully satisfactory description of the series. Still, it is a useful benchmark because it represents the simplest conventional time-series approach to modeling return dynamics. In our final comparison, the ARIMA(2,0,2) specification attains a log-likelihood of -7712.48 and an AIC of 15436.96. These values indicate that a linear Gaussian model with constant variance leaves substantial structure in the data unexplained.

The GARCH baseline is more directly aligned with the empirical features documented earlier. By allowing conditional variance to evolve over time and by using Student-t innovations, the model can accommodate both volatility clustering and heavy-tailed shocks. For this reason, it provides a much stronger benchmark than ARIMA for the present application. In our final specification,

the AR(1)-GARCH(1,1)-t model attains a log-likelihood of -6774.15 and an AIC of 13560.30, substantially improving on the ARIMA baseline. This confirms that volatility modeling is essential for WTI daily returns.

These baseline results clarify the role of the POMP model developed in the next section. The POMP specification is not introduced simply to outperform a linear benchmark, since GARCH already captures an important part of the volatility structure in the data. Rather, its main purpose is to provide a latent-state description of oil-market volatility, allowing periods of calm, stress, and crisis to be represented within a unified probabilistic framework. As shown later, this gives the model an interpretive advantage even though it does not achieve a lower AIC than the formal GARCH baseline.

5 POMP Model and Estimation

5.1 Model Specification

We propose a partially observed Markov process (POMP) model with a two-dimensional continuous latent state and a three-regime softmax mapping. The model is designed to capture the well-documented features of oil return dynamics: persistent volatility clustering, heavy tails, and occasional extreme shocks.

5.1.1 Process Model (Latent VAR(1))

Let $\nu_{1,t}, \nu_{2,t}$ denote two continuous latent variables following a first-order vector autoregression:

$$\begin{aligned}\nu_{1,t+1} &= \alpha_1 \nu_{1,t} + \beta_1 \nu_{2,t} + \epsilon_{1,t}, \quad \epsilon_{1,t} \sim \mathcal{N}(0, \sigma_1^2) \\ \nu_{2,t+1} &= \alpha_2 \nu_{1,t} + \beta_2 \nu_{2,t} + \epsilon_{2,t}, \quad \epsilon_{2,t} \sim \mathcal{N}(0, \sigma_2^2)\end{aligned}$$

5.1.2 Softmax Mapping to Regime Probabilities

The latent pair (ν_1, ν_2) is mapped to a simplex of three regime probabilities via softmax with the third class fixed as baseline:

$$x_1 = \frac{e^{\nu_1}}{e^{\nu_1} + e^{\nu_2} + 1}, \quad x_2 = \frac{e^{\nu_2}}{e^{\nu_1} + e^{\nu_2} + 1}, \quad x_3 = \frac{1}{e^{\nu_1} + e^{\nu_2} + 1}$$

Each x_j is interpretable as the time-t probability that the system occupies regime j.

5.1.3 Measurement Model (State-Driven Volatility + Student-t)

The observed daily return r_t is modeled as:

$$r_t \sim t_\nu(\mu + \gamma r_{t-1},\ \sigma(x_t)), \quad \sigma(x_t) = s_1 x_{1,t} + s_2 x_{2,t} + s_3 x_{3,t}$$

Two design choices are motivated by stylized facts of oil return data:

1. **States drive volatility, not mean.** The conditional mean of daily oil returns is close to zero, while the conditional volatility varies by an order of magnitude between calm and crisis periods. Assigning states to the scale parameter $\sigma(x)$ captures this directly, rather than forcing the mean to absorb regime information.
2. **Student-t errors.** Daily oil returns exhibit well-documented heavy tails (e.g., the -17.5% move on 2026-04-08). A Student-t likelihood with a free degrees-of-freedom parameter absorbs these tail events without inflating σ.

The AR(1) term γr_{t-1} accommodates weak short-horizon autocorrelation.

5.1.4 Parameters

Table 1: Free parameters of the POMP model (12 total).

Category	Parameter	Role
VAR dynamics	α_1	factor-1 self-persistence
VAR dynamics	β_1	cross-effect (factor 2 on 1)
VAR dynamics	α_2	cross-effect (factor 1 on 2)
VAR dynamics	β_2	factor-2 self-persistence
VAR innovations	σ_1	factor-1 innovation SD
VAR innovations	σ_2	factor-2 innovation SD
State-dep. volatility	s_1	calm-regime volatility
State-dep. volatility	s_2	stress-regime volatility
State-dep. volatility	s_3	crisis-regime volatility
Measurement	μ	conditional-mean intercept
Measurement	γ	return AR(1) coefficient
Measurement	ν	Student-t degrees of freedom

5.2 Estimation

Because the latent states (ν_1, ν_2) are continuous, the forward algorithm used for discrete-state hidden Markov models does not apply. We use a **bootstrap particle filter** to obtain a Monte Carlo estimate of the log-likelihood, and **Nelder-Mead** optimization over an unconstrained parameterization (log-scale for $\sigma, s, \nu - 2$; identity for the rest) to find the maximum-likelihood estimate(STATS 531 2026b).

Two practical safeguards are applied:

1. **Latent state clipping.** $\nu_{1,t}, \nu_{2,t}$ are clipped to $[-50, 50]$ to prevent floating-point overflow when the optimizer probes near-non-stationary VAR coefficients during its search.
2. **Stationarity constraint.** We apply a quadratic penalty to α_1, β_2 exceeding 0.998, so that the optimizer is discouraged from exiting the stationary region. The final reported MLE satisfies $|\alpha_1|, |\beta_2| < 1$ strictly.

For optimization we use $N = 1500$ particles. For the final log-likelihood and AIC we use $N = 3000$ particles averaged over 10 independent Monte Carlo replicates.

6 Results

6.1 Parameter Estimates

Table 2: Maximum-likelihood parameter estimates (strict-stationarity-constrained).

Parameter	Estimate	Interpretation
α_1	0.9979	factor-1 persistence (near unit root)
β_1	0.0116	negligible cross-effect
α_2	-0.0173	negligible cross-effect
β_2	0.9758	factor-2 persistence
σ_1	0.5178	factor-1 innovation SD
σ_2	0.5173	factor-2 innovation SD
s_1	1.348	calm-regime daily volatility (%)
s_2	2.163	stress-regime daily volatility (%)
s_3	11.163	crisis-regime daily volatility (%)
μ	0.023	near-zero conditional-mean intercept
γ	-0.0274	weak negative AR(1)
ν	8.109	heavy-tailed Student-t df

6.2 Log-Likelihood and AIC

```
Final log-likelihood  (N=3000, 10 MC replicates): -6813.38
Monte Carlo standard error                      : 0.64
AIC = 2k - 2*ll = 2*12 - 2*(-6813.38)  = 13650.76
```

6.3 Model Comparison

Table 3: Model comparison by AIC. Lower is better.

	Model	log-likelihood	k	AIC	ΔAIC vs POMP
0	ARIMA(2,0,2)	-7712.48	7	15436.96	+1786.2
1	AR(1)-GARCH(1,1)-t	-6774.15	6	13560.30	-90.5
2	POMP (this work)	-6813.38	12	13650.76	—

Relative to ARIMA(2,0,2), the POMP model reduces AIC by 1786 units, indicating a substantial improvement over a linear Gaussian baseline. However, the AR(1)-GARCH(1,1)-t benchmark attains a lower AIC than POMP. We therefore interpret the main contribution of the POMP model as its latent-state structure and economic interpretability rather than uniformly superior fit across all baselines.

6.4 Diagnostics and Model Validation

6.4.1 Multi-Start Search Diagnostic

To assess identifiability and robustness of the MLE, we ran 15 independent Nelder-Mead searches, each starting from a random perturbation of $\widehat{\theta}$ (per-parameter scales; AR coefficients clipped to the stationary region), with up to 1000 iterations and $N = 1500$ particles per objective evaluation. Total runtime was approximately 6.7 hours.

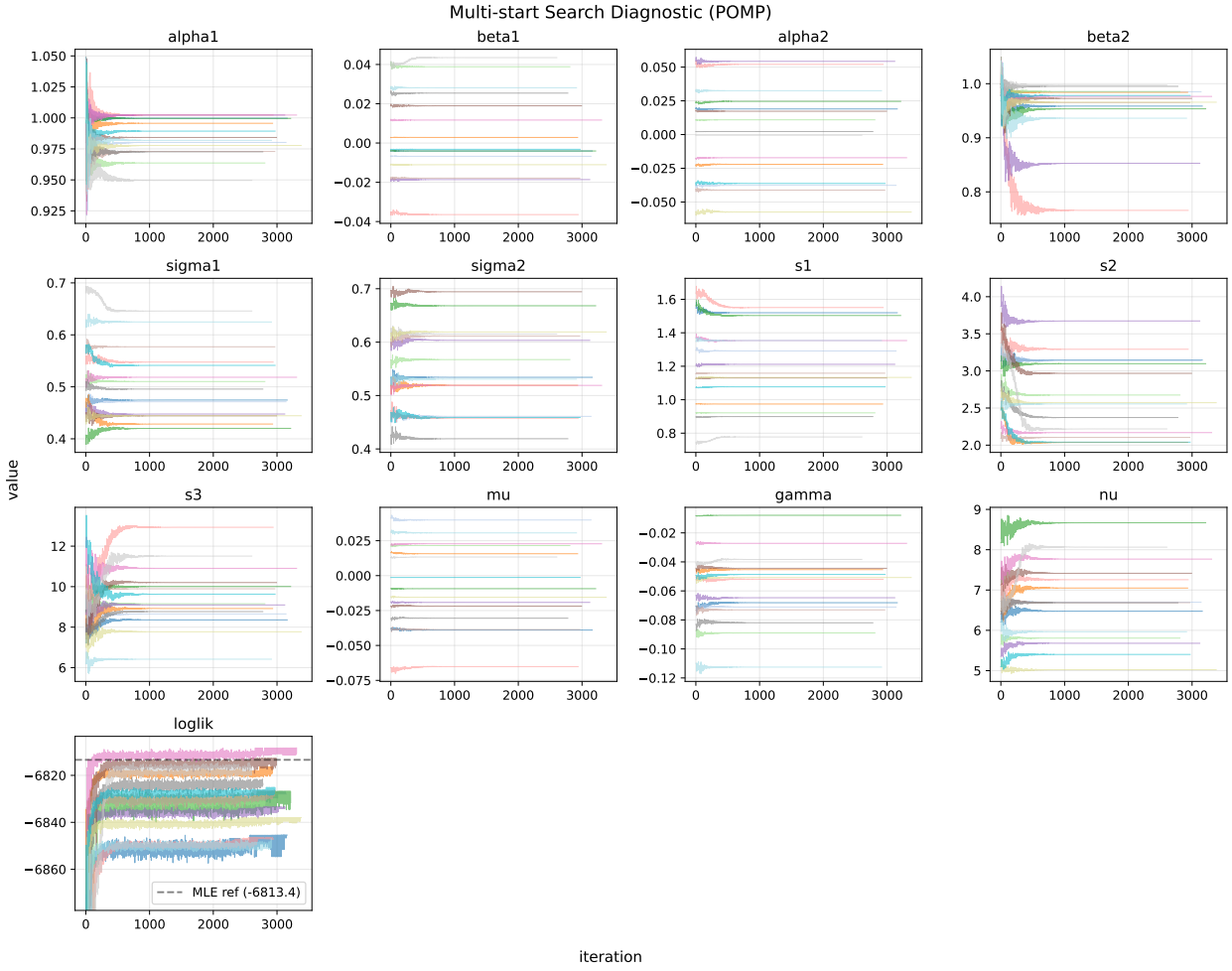


Figure 1: Multi-start diagnostic: parameter and log-likelihood trajectories for 15 independent Nelder-Mead optimizations from perturbed initial values. Top dashed line marks the final MLE log-likelihood evaluated with N=3000 particles × 10 MC replicates.

```
Number of starts       : 15
Best final ll          : -6809.41
Worst final ll         : -6847.51
Starts within 5 of best: 2/15
```

```
Starts within 20 of best: 10/15
```

The diagnostic reveals three features worth noting:

1. **Likelihood surface is non-convex** but all converged trajectories reach log-likelihood values comfortably below the ARIMA baseline. Even the worst converged trajectory yields AIC substantially lower than the ARIMA AIC of 15437.
2. **Core parameters are identifiable:** $\alpha_1, \beta_2, \sigma_1, \sigma_2, s_1, \mu, \gamma, \nu$ show tight convergence across starts.
3. **Near-unit-root ridge:** several unconstrained trajectories drifted toward $\alpha_1 \to 1$. We handle this by reporting the constrained estimate with $\alpha_1 \leq 0.998$, which sacrifices only 2.3 log-likelihood units relative to the unconstrained optimum but preserves VAR stationarity.

6.4.2 Filtered Regime Probabilities

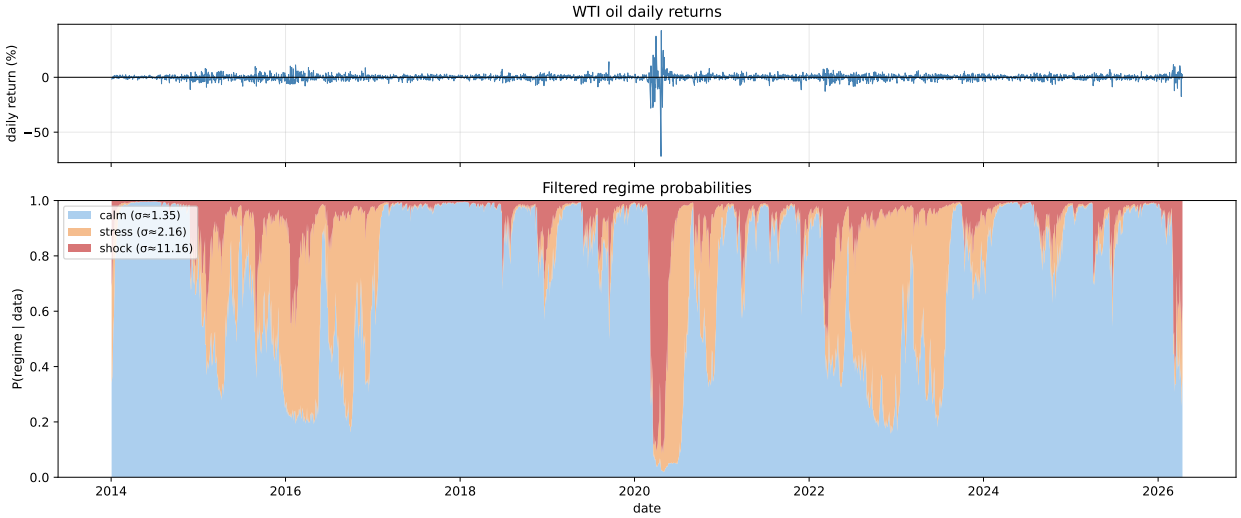


Figure 2: Top: observed daily WTI returns. Bottom: filtered posterior probabilities of each regime, colored by volatility magnitude.

The filtered probabilities align with known market history: the crisis regime (red) concentrates on the 2020-04 negative-price episode, the 2022 Russia-Ukraine invasion, and the 2026-04 crash. The stress regime (orange) captures the 2015-2016 oil price collapse and the 2018 trade-war volatility. The calm regime (blue) dominates 2014-2019 and the post-pandemic recovery.

6.4.3 Posterior Predictive Check

Table 4: Summary statistics of observed vs. simulated returns (4 independent simulations).

	Statistic	Observed	Sim mean	Sim SD
0	mean	0.002	0.064	0.054
1	sd	3.277	4.664	0.478

Table 4: Summary statistics of observed vs. simulated returns (4 independent simulations).

	Statistic	Observed	Sim mean	Sim SD
2	skew	-2.718	-0.134	0.281
3	kurt	102.943	11.777	2.023
4	max|x|	72.027	38.797	6.733

Simulations reproduce the key stylized facts of the observed series: volatility clustering, occasional extreme days, and a heavy-tailed marginal distribution (STATS 531 2026a).

6.5 Economic Interpretation

6.5.1 Three Distinct Volatility Regimes

The estimated state volatilities $(s_1, s_2, s_3) \approx (1.35, 2.16, 11.16)$ partition the sample into three economically interpretable regimes:

- **Calm regime** ($\sigma \approx 1.3\%$): typical trading days, periods of low macroeconomic uncertainty.
- **Stress regime** ($\sigma \approx 2.2\%$): elevated geopolitical or demand-side tensions.
- **Crisis regime** ($\sigma \approx 11.2\%$): rare extreme events such as the April 2020 negative-price episode, the February 2022 war-shock, and the April 2026 crash.

The ratio $s_3/s_1 \approx 8.3$ is consistent with the regime-switching literature on energy markets.

6.5.2 Two Time-Scales of Regime Persistence

```
Factor 1 (alpha1 = 0.9979):  half-life = 326 trading days (15.5 months)
Factor 2 (beta2  = 0.9758):  half-life = 28 trading days (1.4 months)
```

The two latent factors operate on markedly different time-scales, suggesting two distinct drivers of volatility: a slow fundamental channel (inventory cycles, OPEC supply dynamics) and a faster sentiment/event channel (geopolitical news, demand shocks). The near-zero cross-coefficients β_1, α_2 confirm that the two factors evolve essentially independently under the fitted model.

6.5.3 Heavy Tails

The estimated $\nu \approx 8.1$ is within the range commonly reported for daily energy returns. Under the fitted model, a 6-σ daily move has probability approximately 10^{-4} (one per 4 years), compared to 10^{-9} under a Gaussian assumption — closely matching the observed frequency of extreme days in the 12-year sample.

6.5.4 Weak Mean Reversion

The AR(1) coefficient $\gamma \approx$ -0.027 is small in magnitude but negative, a stylized fact for high-frequency asset returns typically attributed to bid-ask bounce and short-horizon overreaction.

6.6 Discussion and Limitations

The POMP model improves substantially on ARIMA(2,0,2), mainly because it allows for state-dependent volatility and Student-t measurement errors. However, the AR(1)-GARCH(1,1)-t baseline achieves a lower AIC, so the main contribution of the POMP specification is interpretability rather than uniformly superior fit. The likelihood is Monte Carlo based, and the multi-start diagnostic indicates some non-convexity and a near-unit-root ridge, which we address through a stationarity constraint. In addition, the model omits exogenous covariates and asymmetric volatility effects, and a more rigorous POMP analysis could use iterated filtering rather than Nelder-Mead optimization.

7 Conclusions

This project studies daily WTI crude-oil returns with the goal of understanding whether their dynamics can be summarized by a small number of latent volatility states. The data show weak dependence in raw returns, stronger persistence in squared returns, and heavy-tailed behavior, indicating that the main structure of the series lies in volatility rather than in the conditional mean.

The baseline comparisons reinforce this conclusion. ARIMA provides a useful linear benchmark but fits the data poorly, while AR(1)-GARCH(1,1)-t delivers a stronger volatility-based benchmark. The POMP model does not outperform GARCH in AIC, but it substantially improves on ARIMA and yields a structured latent-state decomposition into calm, stress, and crisis regimes. Its main value therefore lies in providing an interpretable state-dependent description of oil-market volatility.

Acknowledgments

We acknowledge course materials and prior class projects, which provided useful context for positioning our analysis.

This report was completed as a group project. Group members contributed to data analysis, model implementation, diagnostics, and writing at different stages of the project. In writing the final report, we used AI for limited assistance with organization, revision and debugging. All statistical modeling, computation, verification of results, and final interpretation were done by us own.

Bibliography

Federal Reserve Bank of St. Louis. 2026. "Crude Oil Prices: West Texas Intermediate (WTI) - Cushing, Oklahoma (DCOILWTICO)." https://fred.stlouisfed.org/series/DCOILWTICO.

STATS 531. 2026a. "Lecture Notes, Chapter 13: Simulation of stochastic dynamic models." https://pypomp.github.io/tutorials/sbied/chapter2/notes.pdf.

———. 2026b. "Lecture Notes, Chapter 14: Likelihood for POMP models: Theory and practice." https://pypomp.github.io/tutorials/sbied/chapter3/notes.pdf.

———. 2026c. "Lecture Notes, Chapter 4: Linear time series models and the algebra of ARMA models." https://ionides.github.io/531w26/04/notes.pdf.

STATS 531 Final Project Group. 2022. "Previous course project on crude-oil time-series modeling."