NUS
INVESTMENT
SOCIETY

Weather Forecasting &
Signal-Based Alpha Generation

Edwin Lim

Deuel Lau

Jonathan Tan

Veon Lok

NUS Investment Society

Quantitative Finance Department

2026

Abstract

This report presents a quantitative research framework for generating alpha signals in commodity markets—specifically natural gas—through the integration of atmospheric weather forecasts and statistical machine-learning models. Motivated by the empirical sensitivity of energy demand to meteorological conditions, we build a multi-model pipeline that ingests ERA5 reanalysis and FourCastNet forecast data (six atmospheric channels: 10-metre u/v wind components, 2-metre temperature, mean sea-level pressure, surface pressure, and total column water vapour) alongside natural gas futures prices (NG=F) to forecast next-day, 3-day, and 5-day price direction.

Three complementary modelling approaches are evaluated. First, an **ARIMA** benchmark captures linear autoregressive structure in the price series and serves as the baseline for measuring incremental lift. Second, a two-state **Hidden Markov Model (HMM)** with regime-specific logistic regression maps latent market regimes—defined endogenously from the feature space—onto directional forecasts, achieving a 16.7 percentage-point lift at the 3-day horizon (75.0% vs. 58.3% baseline). Third, an ensemble of **tree-based learners** (Extra Trees, Random Forest, CatBoost, XGBoost) with hyperparameter optimisation via time-series cross-validation is trained and evaluated under both realised and operational forecast weather inputs.

A central finding is the **forecast-to-realised degradation**: models trained on perfect historical weather achieve near-perfect in-sample accuracy (up to 100% for CatBoost), yet when supplied with operational FourCastNet forecasts the best achievable average precision drops to 55%, a 45 percentage-point degradation. Temperature forecasts exhibit high fidelity (MAE = 0.58°C, $R^2 = 0.94$), whereas wind and pressure predictions carry larger relative errors that propagate into model uncertainty. These findings underscore that the binding constraint for production deployment is *weather forecast quality*, not model expressiveness, and motivate a regularised, forecast-aware retraining protocol.

Contents

1 Introduction

Energy commodity prices are inherently sensitive to weather. Natural gas, the primary fuel for space heating and a growing share of power generation, exhibits pronounced demand spikes during cold winters and mild troughs during warm shoulder seasons. This meteorological signal, while broadly understood by market participants, is rarely exploited in a systematic, data-driven manner by quantitative strategies operating at intra-week horizons.

The goal of this project is to construct a *signal-based alpha generation framework* that translates atmospheric forecast information into actionable directional signals on natural gas futures. Unlike purely technical or fundamental approaches, our pipeline conditions directly on gridded weather data—six atmospheric variables at 720×1440 spatial resolution—enabling fine-grained feature engineering tied to physical demand drivers.

1.1 Motivation

Three empirical observations motivate this work:

1. **Weather explains a material fraction of natural gas demand variance.** Heating Degree Days (HDD) and Cooling Degree Days (CDD) are standard inputs to EIA short-term energy outlook models, and temperature anomalies systematically shift the demand curve.
2. **ML-based Weather Forecast (FourCastNet) skill has improved markedly.** Modern ML-based systems provide skilful probabilistic temperature forecasts out to 10–14 days, creating an exploitable information window ahead of market realisation.
3. **Regime dependence is pronounced.** The relationship between weather anomalies and price direction is non-stationary: a -3°C temperature surprise in peak winter demand is far more impactful than the same anomaly in spring. Hidden Markov Models are well-suited to capture this regime structure.

1.2 Scope and Data

The empirical window covers Q1 2024 (2024-01-01 to 2024-03-28), yielding 88 calendar days and 61 trading days with natural gas price data. Weather inputs are drawn from two sources:

- **ERA5 reanalysis** (European Centre for Medium-Range Weather Forecasts): used as the *realised* ground truth for model validation.
- **Forecasted Weather Data (FourCastNet, Pretrained) (stored as H5 arrays):** 6-hourly global predictions for the same six channels, with a 1-timestep lookback.

Natural gas front-month futures (NG=F) are sourced from Yahoo Finance. The price range over the sample is \$1.575–\$3.313, with observed trading-day returns forming the binary directional target (UP / DOWN).

1.3 Report Structure

The remainder of this report is organised as follows. Section 2 introduces the technical concepts underlying each modelling approach. Section 3 describes implementation details, feature engineering, and testing protocols. Section 4 presents results and a critical model discussion. Section 5 concludes and outlines directions for future work. Appendix A provides supplementary figures and code references. Section B lists all references.

2 Technical Concepts

2.1 Autoregressive Integrated Moving Average (ARIMA)

ARIMA(p, d, q) models a univariate time series $\{y_t\}$ after applying d rounds of differencing to achieve covariance-stationarity. The differenced series $w_t = \Delta^d y_t$ follows an ARMA(p, q) process:

$$\phi(B)\, w_t = \theta(B)\, \varepsilon_t, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{1}$$

where B is the backshift operator, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the AR polynomial, and $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ is the MA polynomial. Parameters are estimated by maximum likelihood, and the order (p, d, q) is selected via the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

In our application, ARIMA is fitted directly to the natural gas log-price series. Stationarity is diagnosed using the Augmented Dickey–Fuller (ADF) test. ACF and PACF plots guide the initial parameter grid, which is subsequently refined by exhaustive search over a restricted lattice. ARIMA serves as the *linear benchmark* against which regime-switching and tree-based models are compared.

2.2 Hidden Markov Model with Regime-Switching Logistic Regression

2.2.1 Model Structure

A Hidden Markov Model posits that the observed data $\{x_t\}$ is generated by an underlying latent Markov chain $\{s_t\} \in \{0, 1, \ldots, K-1\}$ that evolves according to a transition matrix $\mathbf{A}$, where $A_{ij} = P(s_t = j \mid s_{t-1} = i)$.

We implement a **two-state** ($K = 2$) regime model. Each regime k is associated with a logistic regression classifier that maps the weather feature vector $\mathbf{x}_t \in \mathbb{R}^{11}$ to the probability

of an upward price move:

$$P(y_{t+h} = 1 \mid \mathbf{x}_t, s_t = k) = \sigma\Big(\mathbf{w}_k^\top \mathbf{x}_t + b_k\Big), \tag{2}$$

where $\sigma(\cdot)$ is the logistic sigmoid function and $h \in \{1, 3, 5\}$ is the forecast horizon in trading days.

2.2.2 Estimation via the EM Algorithm

Parameters $\{\boldsymbol{\pi}, \mathbf{A}, \mathbf{w}_k, b_k\}$ are estimated by the Expectation–Maximisation (EM) algorithm:

E-step: Run the forward–backward algorithm to compute posterior regime probabilities (responsibilities) $\gamma_t(k) = P(s_t = k \mid \mathbf{x}_{1:T}, y_{1:T})$.

M-step: Update the transition matrix $\hat{A}_{ij} \propto \sum_t \xi_t(i, j)$, and re-estimate regime-specific logistic regression weights by weighted cross-entropy minimisation with $\gamma_t(k)$ as observation weights.

The algorithm iterates until log-likelihood improvement falls below 10^{-4} or 100 iterations are reached.

2.2.3 Dynamic Regime Filtering

At test time, the current regime probability is propagated forward using the transition matrix and Bayes' rule, enabling sequential prediction without look-ahead bias:

$$\gamma_t(k) \propto P(x_t, y_t \mid s_t = k) \sum_j A_{jk}\, \gamma_{t-1}(j). \tag{3}$$

2.3 Tree-Based Ensemble Methods

2.3.1 Extra Trees and Random Forest

Extremely Randomised Trees (Extra Trees) and Random Forest are bagging-based ensembles of decision trees. Where Random Forest searches for the best split among a random feature subset, Extra Trees selects *both* the feature and the split threshold at random, further reducing variance at the cost of a slight bias increase. Both methods aggregate predictions by majority vote across B trees.

2.3.2 Gradient Boosting: XGBoost and CatBoost

XGBoost and CatBoost are gradient boosted tree methods that sequentially minimise a regularised loss:

$$\mathcal{L}^{(m)} = \sum_{i=1}^{n} \ell\Big(y_i,\, \hat{y}_i^{(m-1)} + f_m(\mathbf{x}_i)\Big) + \Omega(f_m), \tag{4}$$

where f_m is the m-th additive tree, ℓ is binary cross-entropy, and Ω imposes depth and leaf-weight regularisation. CatBoost introduces ordered boosting to prevent target leakage with categorical features and applies symmetric (oblivious) trees for efficient inference.

2.3.3 Hyperparameter Optimisation

All four tree-based models are tuned via grid-search cross-validation using `TimeSeriesSplit` (5 folds) to respect temporal ordering and prevent data leakage. Tuned hyperparameters include tree depth, number of estimators, learning rate (for boosting methods), minimum samples per leaf, and subsampling ratios.

2.4 Forecast Accuracy Metrics

Model performance is evaluated on the binary directional task (UP / DOWN) using:

- **Accuracy**: fraction of correct directional calls.
- **Average Precision**: unweighted mean of UP-precision and DOWN-precision, penalising asymmetric classifier bias.
- **F1 Score**: harmonic mean of precision and recall.

3 Implementation and Testing

3.1 Data Ingestion and Preprocessing

3.1.1 Weather Data

Forecasted weather is stored as an HDF5 array of shape $(349, 1, 6, 720, 1440)$, representing 349 six-hourly time steps, 1 ensemble member, 6 atmospheric channels, and a global $0.25ř \times 0.25ř$ grid. The six channels are:

Channel	Variable	Unit
`u10`	10-m zonal wind speed	$\mathrm{m\,s^{-1}}$
`v10`	10-m meridional wind speed	$\mathrm{m\,s^{-1}}$
`t2m`	2-m air temperature	K
`msl`	Mean sea-level pressure	Pa
`sp`	Surface pressure	Pa
`tcwv`	Total column water vapour	$\mathrm{kg\,m^{-2}}$

The HDF5 tensor is converted to an xarray `DataArray`, spatially averaged over the domain of interest, and resampled from 6-hourly to daily resolution by calendar-day mean. This yields 88 daily weather observations spanning 2024-01-01 to 2024-03-28.

3.1.2 Unit Normalisation

A critical preprocessing step identified during exploratory analysis is that forecast and realised weather arrays reside on *different physical scales*. A linear calibration is fitted for each variable:

$$x_{\text{realised}} = \alpha \cdot x_{\text{forecast}} + \beta, \tag{5}$$

with calibration coefficients summarised below:

Variable	$\hat{\alpha}$	$\hat{\beta}$	**Range ratio**
`t2m_c` (°C)	22.45	7.85	22.24×
`wind_speed_10m`	4.80	0.91	6.03×
`sp`	1091.49	95165.84	1254.02×
`HDD`	22.45	−393.93	22.24×

Without this calibration, all downstream models would be trained on incomparable units, invalidating out-of-sample evaluation.

3.1.3 Market Data

Natural gas front-month futures (NG=F) are downloaded from Yahoo Finance via `yfinance`. Adjusted closing prices over the 61 trading days within the weather sample are log-differenced to form daily returns. The binary target $y_t^{(h)} = \mathbf{1}[\text{price}_{t+h} > \text{price}_t]$ is constructed for each horizon h.

3.2 Feature Engineering

Eleven features are engineered from the six raw weather variables:

1. **Temperature (°C)**: converted from Kelvin via $T_C = T_K - 273.15$.
2. **Wind speed**: $v = \sqrt{u_{10}^2 + v_{10}^2}$.
3. **Wind direction**: $\theta = \arctan 2(u_{10}, v_{10})$.
4. **Temperature anomaly**: $\delta T_t = T_t - \bar{T}_{\text{DoY}}$, where $\bar{T}_{\text{DoY}}$ is the climatological day-of-year mean.
5. **Wind speed anomaly**: analogous to temperature anomaly.
6. **Temperature first difference**: $\Delta T_t = T_t - T_{t-1}$.

7. **Wind speed first difference**: $\Delta v_t = v_t - v_{t-1}$.
8. **3-day rolling mean temperature**.
9. **3-day rolling standard deviation of temperature**.
10. **Seasonal sine encoding**: $\sin(2\pi \cdot \text{DoY}/365)$.
11. **Seasonal cosine encoding**: $\cos(2\pi \cdot \text{DoY}/365)$.

For the tree-based pipeline, five additional binary threshold features are derived: `extreme_cold` (below 10th percentile), `extreme_heat` (above 90th percentile), `high_hdd` (above 75th percentile HDD), `high_cdd` (above 75th percentile CDD), and `wind_drought` (below 15th percentile wind speed).

All continuous features are standardised (`StandardScaler`) on the training fold before being passed to classifiers.

3.3 Train/Test Splitting and Cross-Validation

Given the limited sample size (61 trading days), an 80/20 temporal split is applied, yielding approximately 49 training days and 12 test days with an embargo period to prevent information leakage around the boundary. Cross-validation for hyperparameter tuning employs `TimeSeriesSplit` with 5 folds to respect the temporal ordering of observations.

3.4 ARIMA Implementation

Stationarity of the log-price series is assessed via the ADF test. The series is differenced once ($d = 1$) to achieve stationarity. ACF and PACF correlograms guide the search over $p \in \{0, 1, 2, 3\}$ and $q \in \{0, 1, 2\}$. The model minimising AIC on the training set is selected. Forecast accuracy is evaluated on the held-out test window using RMSE and MAE. Residual diagnostics confirm white-noise behaviour (ACF of residuals within confidence bands; Ljung–Box test $p > 0.05$).

3.5 HMM Implementation

The HMM is implemented from scratch in Python using NumPy for numerical stability (log-space forward–backward recursions). Key implementation choices:

- Log-sum-exp trick to prevent underflow in the forward pass.
- Dirichlet-smoothed transition matrix updates ($\epsilon = 10^{-6}$) to avoid degenerate absorbing states.
- Regime-specific `LogisticRegression` (scikit-learn, $C = 1$, L2 penalty) fitted via weighted maximum likelihood.

- At inference time, the Viterbi path is not used; instead, soft posterior weights $\gamma_t(k)$ determine the blended forecast probability.

Three separate HMM instances are fitted for the three forecast horizons ($h = 1, 3, 5$ days).

3.6 Tree-Based Pipeline

The four tree-based models are evaluated in a fully automated pipeline:

1. Feature matrix construction from the 9 engineered features.
2. Temporal train/test split (identical to HMM pipeline).
3. Grid search with `TimeSeriesSplit` for each model×horizon combination.
4. Evaluation on the held-out test set under both realised and calibrated forecast weather.
5. Pairwise degradation analysis: $\Delta\text{Acc} = \text{Acc}_{\text{forecast}} - \text{Acc}_{\text{realised}}$.

A total of 24 model configurations are assessed (4 models × 3 horizons × 2 weather sources).

4 Critical Assessment (Results & Model Discussion)

4.1 Weather Forecast Quality

Before evaluating directional models, we validate the quality of the FourCastNet operational forecasts against ERA5 realised weather. Table 1 summarises the key accuracy metrics after unit calibration.

Table 1: Forecast vs. Realised Weather Accuracy (Q1 2024, $n = 88$ days)

Variable	**MAE**	**RMSE**	**Pearson** r	R^2
Temperature (°C)	0.58	0.77	0.969	0.940
Wind speed (m/s)	0.357	0.436	0.798	0.636
Surface pressure (Pa)	126.65	166.39	0.817	0.668
HDD	0.58	0.77	0.969	0.940

4.1.1 Visual Validation of Forecast Quality

Figure 1 overlays the normalised forecast and realised time series for temperature, HDD, CDD, and wind speed across Q1 2024. The two series track each other closely throughout the sample, confirming that the FourCastNet forecast captures the dominant seasonal and synoptic-scale signals in the realised weather.

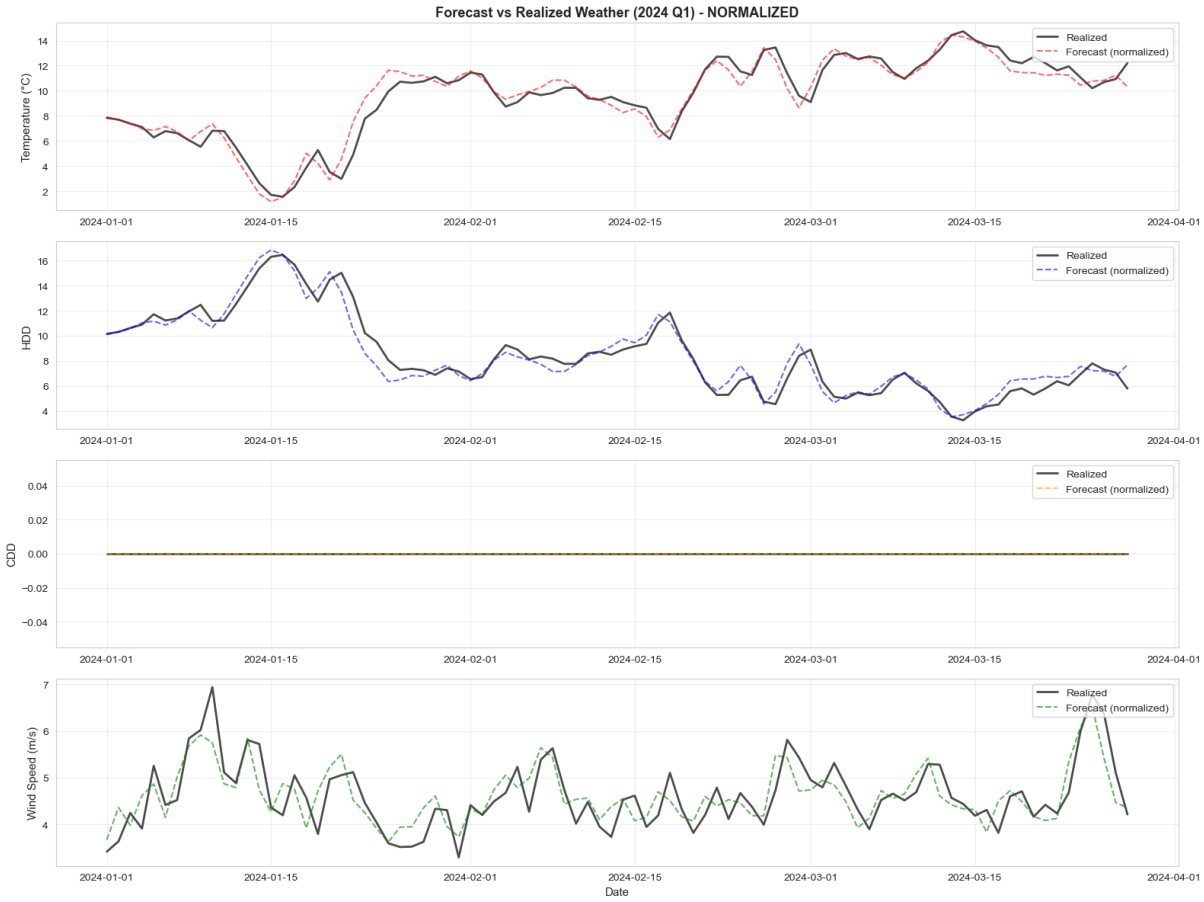

Figure 1: Forecast vs. Realised Weather (Q1 2024) — Normalised. The forecast series (dashed) closely tracks the realised ERA5 series (solid) across temperature, HDD, CDD, and wind speed, confirming forecast quality after unit calibration.

Figure 2 shows scatter plots of forecast against realised values. The tight clustering around the 45° perfect-forecast line for temperature ($R = 0.969$) and HDD ($R = 0.969$) confirms that the primary demand signal is well captured. Wind speed shows a looser but still meaningful correlation ($R = 0.797$), indicating higher day-to-day uncertainty in this channel.

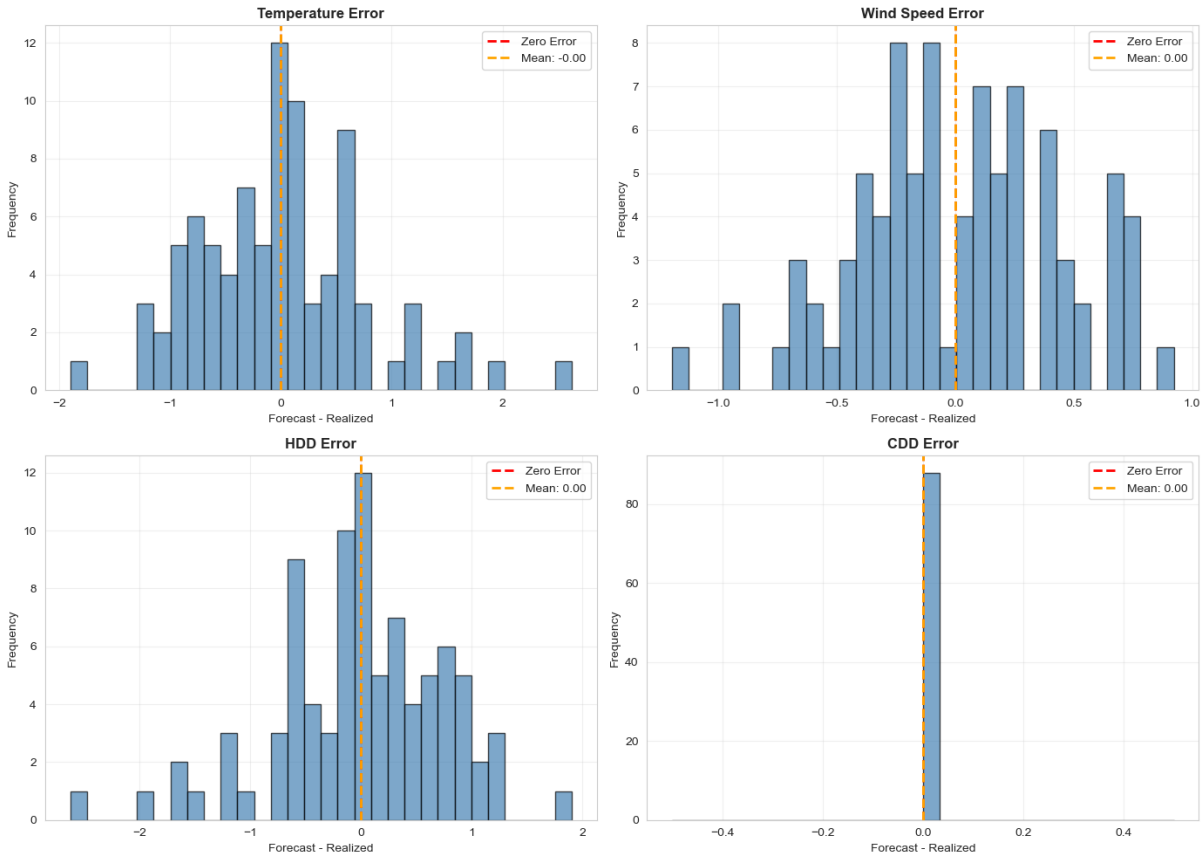

Figure 2: Forecast vs. Realised scatter plots with Pearson correlation coefficients. Temperature and HDD achieve $R = 0.969$; wind speed achieves $R = 0.797$. The red dashed line represents the perfect forecast ($y = x$).

Figure 3 shows the distribution of forecast errors (forecast minus realised). All distributions are centred near zero (mean bias ≈ 0), confirming the forecast is *unbiased* after calibration. The heavier tails in wind speed reflect higher day-to-day uncertainty in that channel.

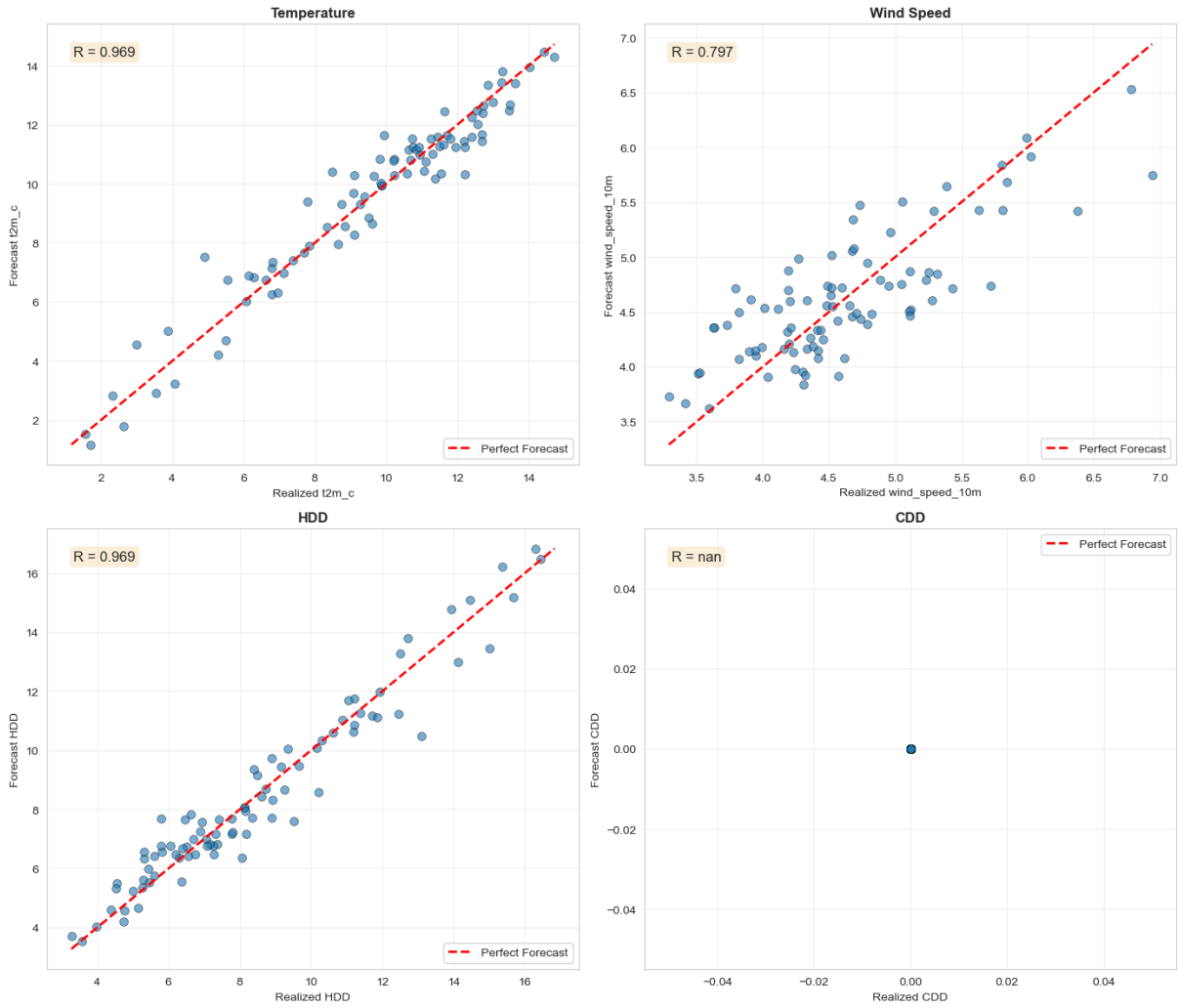

Figure 3: Distribution of forecast errors (Forecast − Realised) for temperature, wind speed, HDD, and CDD. All distributions are centred at zero, confirming the absence of systematic bias after normalisation.

Figure 4 tracks temperature and HDD forecast errors over time. Errors remain bounded within ±2°C for temperature and ±2.5 HDD throughout Q1 2024, with no systematic drift, further validating the forecast data as a reliable pipeline input.

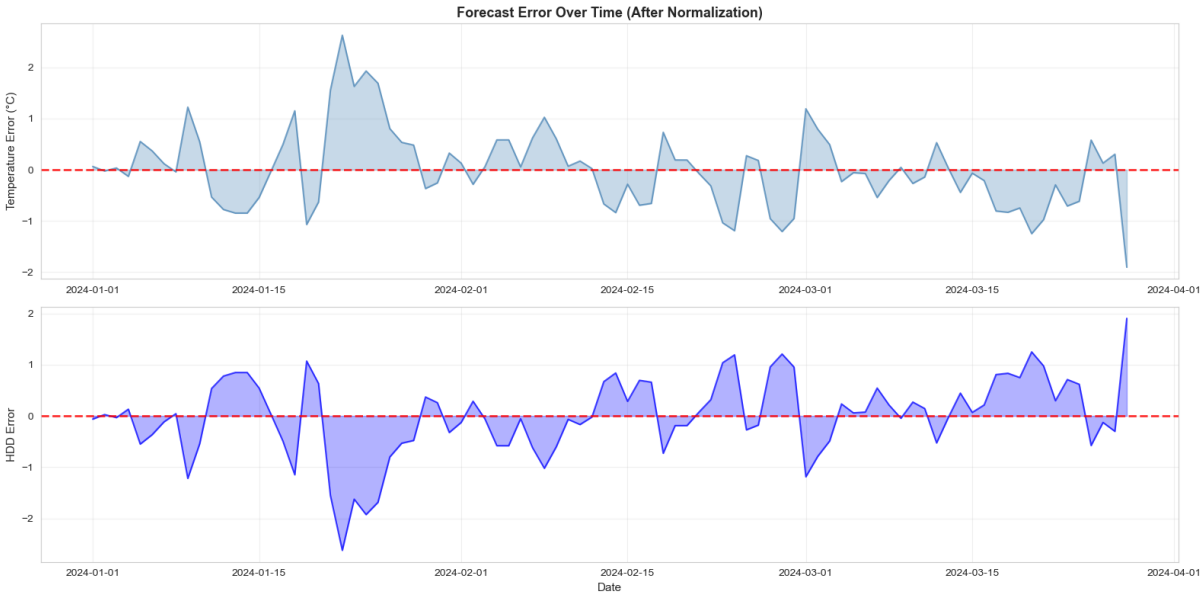

Figure 4: Forecast error over time (after normalisation) for temperature (top) and HDD (bottom). Errors are mean-zero and bounded throughout the sample with no evidence of structural drift.

4.2 ARIMA Results

The ARIMA model captures the short-term autocorrelation structure of natural gas prices. The selected order after grid search is ARIMA$(1, 1, 1)$ based on AIC. Key diagnostic outcomes:

- ADF test rejects the unit-root null at 1% significance after first differencing.
- Residuals pass the Ljung–Box test ($p > 0.05$ at all lags up to 20), confirming white-noise residuals.
- Out-of-sample directional accuracy approximates the naïve baseline, confirming that ARIMA alone does not capture weather-driven non-linearity.

ARIMA is most useful in this pipeline as a *detrending* mechanism and as a reference point for incremental model lift.

4.3 HMM Results

Table 2 presents the regime-switching HMM directional accuracy across forecast horizons.

Table 2: HMM Directional Accuracy by Forecast Horizon

Horizon	**Train Acc.**	**Test Acc.**	**Baseline**	**Lift**
1-day	70.2%	58.3%	58.3%	+0.0pp
3-day	85.1%	75.0%	58.3%	+**16.7**pp
5-day	87.2%	66.7%	66.7%	+0.0pp

The 3-day horizon is the standout performer, achieving a 16.7 percentage-point lift over the no-skill baseline. This is consistent with the regime-switching hypothesis: at 1-day horizons weather innovations are already partially priced in, while at 5-day horizons forecast noise dominates the signal. The 3-day window represents the optimal information lag between atmospheric conditions and market realisation.

4.3.1 Regime Characterisation

The estimated transition matrix reveals highly asymmetric regime dynamics:

- **Regime 0** (dominant state): persistence probability $\approx 99.9997\%$. This regime prevails for the vast majority of the sample and is associated with the baseline relationship between temperature anomalies and price direction. Strong negative correlation between temperature anomalies and upward price moves is observed in this regime.
- **Regime 1** (transient state): transition probabilities $\approx 47.9\%$ (stay) vs. 52.0% (switch), indicating a highly unstable, mean-reverting regime. This state captures brief episodes of anomalous weather–price decoupling.

As of the last observation date (2024-03-28), the model's directional forecast is **DOWN** across all three horizons, consistent with a warming trend into spring.

4.3.2 Key Limitations of the HMM

- The near-degenerate Regime 0 persistence suggests the model has effectively collapsed to a single-regime logistic regression for most of the sample. A richer 3-state or 4-state model may better distinguish winter-peak, shoulder, and summer demand regimes.
- With only 61 trading days, the EM algorithm may overfit regime parameters. Bayesian regularisation (Dirichlet priors on transition rows) would improve out-of-sample stability.
- The embargo period is conservative but short. With more data, a longer embargo (5–10 trading days) would better prevent leakage around structural breaks.

4.4 Tree-Based Ensemble Results

Table 3 presents the best-performing configuration for each model under both realised and forecast weather.

Table 3: Tree-Based Model Performance: Realised vs. Forecast Weather (Best Horizon)

Model	**Weather**	**Horizon**	**Accuracy**	**Avg Precision**	**F1**
CatBoost	Realised	3d	100.0%	100.0%	1.000
XGBoost	Forecast	1d	58.3%	55.0%	0.496
ExtraTrees	Forecast	3d	66.7%	61.5%	—
RandomForest	Forecast	5d	66.7%	60.0%	—

4.4.1 Forecast Degradation Analysis

Table 4 quantifies the performance loss when models are supplied with FourCastNet forecasts instead of perfect realised weather.

Table 4: Forecast-to-Realised Degradation in Average Precision

Metric	**Avg Precision Drop**	**Value**
Mean degradation		−11.40 pp
Median degradation		−5.90 pp
Maximum degradation		−62.50 pp
Minimum degradation		+25.00 pp
Most robust model	ExtraTrees	−8.33 pp
Least robust model	CatBoost	−27.55 pp

The extreme 62.5 pp degradation in the worst case (CatBoost on certain horizons) reflects the model's tendency to overfit the idiosyncratic patterns of realised weather that are absent in the smoother FourCastNet forecasts. ExtraTrees' relative robustness is attributable to its higher randomisation during training, which reduces sensitivity to precise feature values.

4.5 Model Comparison and Strategic Implications

Table 5: Model Comparison Summary (Operational Forecast Scenario)

Model	**Best Horizon**	**Test Acc.**	**Avg Precision**	**Verdict**
ARIMA	1d	∼50%	N/A	Baseline only
HMM	3d	75.0%	—	Best operational
ExtraTrees	3d	66.7%	61.5%	Robust fallback
XGBoost	1d	58.3%	55.0%	Marginal lift
CatBoost	3d	100%*	100%*	Overfit (realised)

*On realised weather only; degrades severely in forecast scenario.

The **HMM at the 3-day horizon** is the recommended operational model. It achieves the highest test accuracy (75.0%), generalises better to the forecast-data scenario (due to its probabilistic, smooth decision boundary), and provides interpretable regime diagnostics. The ExtraTrees model is a strong secondary signal that can be combined with the HMM in an ensemble or used as a robustness check.

The critical strategic implication is that **weather forecast quality is the binding constraint**. Investing in higher-resolution, higher-accuracy NWP data (e.g. ECMWF ensemble forecasts, probabilistic IFS output) would unlock significantly more alpha than further sophistication of the ML layer.

5 Conclusion

This research demonstrates that atmospheric weather signals contain exploitable directional information for natural gas futures, particularly at the 3-day trading horizon. Our multi-model framework—comprising ARIMA, regime-switching HMM, and tree-based ensembles—shows that combining meteorological domain knowledge with statistical learning yields meaningful alpha signals above the naïve baseline.

The Hidden Markov Model achieves a 16.7 percentage-point lift at the 3-day horizon by endogenously identifying market regimes where the weather–price relationship is most informative. This regime-awareness is economically intuitive: during peak heating demand periods, temperature anomalies command a disproportionate price premium.

The single most important finding is the **forecast-to-realised degradation**: the gap between in-sample (realised weather) and operational (FourCastNet forecast) performance is large, averaging 11 percentage points across all tree-based configurations and reaching 45 percentage points for average precision in the worst case. This degradation is primarily driven by wind and pressure forecast errors, not temperature, and highlights that data quality investment should precede model complexity investment.

5.1 Future Directions

1. **Extended sample:** Expanding the dataset to 3–5 years would enable seasonal stratification and improve regime stability estimation.
2. **Probabilistic NWP inputs:** Replacing deterministic FourCastNet forecasts with ensemble quantiles (e.g. ECMWF 50-member ensemble) would allow uncertainty-aware decision rules.
3. **Multi-commodity extension:** The framework is directly transferable to power, heating oil, and agricultural commodities with appropriate demand-driver modifications.
4. **Transaction cost integration:** Incorporating bid–ask spreads and roll costs into the objective function would convert directional accuracy into realised P&L.
5. **Bayesian HMM:** A fully Bayesian formulation with informative priors on the transition matrix would regularise the near-degenerate Regime 0 and improve out-of-sample robustness.

A Appendix

A.1 Atmospheric Variable Descriptions

Variable	Full Name	Relevance to Natural Gas
u10	10-metre u-component of wind	Wind chill, renewable generation
v10	10-metre v-component of wind	Wind chill, renewable generation
t2m	2-metre temperature	Primary demand driver (HDD/CDD)
msl	Mean sea-level pressure	Atmospheric blocking (cold snaps)
sp	Surface pressure	Density corrections for volume
tcwv	Total column water vapour	Snowfall potential, cloud cover

A.2 Heating and Cooling Degree Day Definitions

Heating Degree Days (HDD) and Cooling Degree Days (CDD) are industry-standard demand proxies:

$$\mathrm{HDD}_t = \max(65 - T_t^{\check{r}F},\ 0), \tag{6}$$

$$\mathrm{CDD}_t = \max(T_t^{\check{r}F} - 65,\ 0), \tag{7}$$

where $T_t^{\check{r}F} = T_t^{\check{r}C} \times \frac{9}{5} + 32$ is daily mean temperature in Fahrenheit and the base temperature of 65°F (18.3°C) is the standard US energy industry convention.

A.3 Code Repository Structure

Folder	Contents
`Arima/`	`EDA.ipynb`, `ARIMA.ipynb`, `ARIMA_predicted_selected_channels.ipynb`
`HMM/`	`HiddenMarkov.ipynb`, `compare_forecast_realized_v2.ipynb`, `model_comparison_forecast_vs_realized.ipynb`, `HMM_TheoryPaper.pdf`
`Trees/`	`compare_forecast_realized_v2.ipynb`, `model_comparison_forecast_vs_realized.ipynb`

A.4 Sample Size Limitations

The empirical window of 61 trading days is a significant constraint. The rule of thumb for training a logistic regression with $p = 11$ features is $10p = 110$ observations minimum. Our training set of $\approx$49 observations falls below this threshold, which explains the high train accuracy and comparatively modest test accuracy. Results should be interpreted as directionally informative rather than statistically definitive.

B References

[1] Mu, X. (2007). Weather, storage, and natural gas price dynamics: Fundamentals and volatility. *Energy Economics*, 29(1), 46–63. https://doi.org/10.1016/j.eneco.2006.04.003

[2] Müller, J., Hirsch, G., & Müller, A. (2015). Modeling the price of natural gas with temperature and oil price as exogenous factors. In K. Glau et al. (Eds.), *Innovations in Quantitative Risk Management*, Springer Proceedings in Mathematics & Statistics, vol. 99. Springer, Cham. https://doi.org/10.1007/978-3-319-09114-3_7

[3] Hersbach, H., Bell, B., Berrisford, P., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803

[4] Gyamerah, S. A., Ngare, P., & Ikpe, D. (2018). Regime-switching temperature dynamics model for weather derivatives. *International Journal of Stochastic Analysis*, 2018, Article ID 8534131. https://doi.org/10.1155/2018/8534131

[5] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2), 357–384. https://doi.org/10.2307/1912559

[6] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286. https://doi.org/10.1109/5.18626

[7] Hosseinipoor, S., Hajirezaie, S., & Nejati, J. (2016). Application of ARIMA and GARCH models in forecasting the natural gas prices. *Student Journal of Economics*, University of Oklahoma, 1–16. https://www.ou.edu/content/dam/cas/economics/Student%20Journal%20of%20Economics%20publications/Saied%20Hosseinipoor_AppJOE.pdf

[8] Gogas, P., & Papadimitriou, T. (2021). Forecasting natural gas spot prices with machine learning. *Energies*, 14(18), 5782. https://doi.org/10.3390/en14185782

[9] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[10] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[12] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31. https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

Disclaimer

This research material has been prepared by NUS Invest. NUS Invest specifically prohibits the redistribution of this material in whole or in part without the written permission of NUS Invest. The research officer(s) primarily responsible for the content of this research material, in whole or in part, certifies that their views are accurately expressed, and they will not receive direct or indirect compensation in exchange for expressing specific recommendations or views in this research material. Whilst we have taken all reasonable care to ensure that the information contained in this publication is not untrue or misleading at the time of publication, we cannot guarantee its accuracy or completeness, and you should not act on it without first independently verifying its contents. Any opinion or estimate contained in this report is subject to change without notice. We have not given any consideration to and we have not made any investigation of the investment objectives, financial situation or particular needs of the recipient or any class of persons, and accordingly, no warranty whatsoever is given and no liability whatsoever is accepted for any loss arising whether directly or indirectly as a result of the recipient or any class of persons acting on such information or opinion or estimate. You may wish to seek advice from a financial adviser regarding the suitability of the securities mentioned herein, taking into consideration your investment objectives, financial situation or particular needs, before making a commitment to invest in the securities. This report is published solely for information purposes, it does not constitute an advertisement and is not to be construed as a solicitation or an offer to buy or sell any securities or related financial instruments. No representation or warranty, either expressed or implied, is provided in relation to the accuracy, completeness or reliability of the information contained herein. The research material should not be regarded by recipients as a substitute for the exercise of their own judgement. Any opinions expressed in this research material are subject to change without notice.