February 9, 2026

A Simple Forward Rate Analysis

Iason Baldes

iasonbaldes@gmail.com
Sydney, Australia

Abstract

Constant volatility of forward rates is a key assumption underlying common benchmark pricing formulas for interest rate derivatives such as caps/floors and swaptions. We analyse the drift and volatility of a simple forward rate using ten years of federal reserve data with daily sampling. Empirically, we find a drift consistent with zero and fluctuations incompatible either with normal or log normal distributions of constant volatility, as is widely acknowledged in the literature. Using a GARCH analysis, we show the data exhibit volatility clustering. We train neural networks of various architectures to perform volatility forecasting over three month timeframes and compare to the benchmark GARCH forecast. Quantitatively, we find small improvements on the GARCH benchmark with the dataset used here, and with interesting qualitative differences in the predicted time evolution of the volatility.

Contents

1 Introduction and Motivation

The zero coupon bond (ZCB) price is given by

$$P(t, T) = e^{-y(t,T)\,T}, \tag{1}$$

where t is the current time, T is the maturity, and $y(t, T)$ is the yield.

Consider now the simple forward rate (SFR) observed at time t. This is the non-compounding interest rate for borrowing between T_0 and T_1. The zero arbitrage condition — in a liquid market in which participants can borrow and lend at the rates $y(t, T)$ — gives for the forward rate

$$F(t, T_0, T_1) = \frac{1}{\delta}\left(\frac{P(t, T_0)}{P(t, T_1)} - 1\right), \tag{2}$$

where $\delta = T_1 - T_0$. To recover this rate from no arbitrage arguments, note a participant in the market can replicate the instrument by borrowing money at time t of some notional amount until T_1, and immediately issuing a loan with the same notional amount to maturity T_0. Departures from the above forward rate would then allow arbitrage opportunities. (Forward rate agreements typically involve one party paying fixed and the other paying the floating rate, the fixed rate is the same as the above. Moreover, we limit ourselves to US government bond data, while for commercial agreements other benchmarks such as SOFR will be relevant.)

Interestingly, the forward rate is fixed by the non-arbitrage condition independently of any modelling about how we expect $P(t, T_0)$ and $P(t, T_1)$ to evolve with time. To price derivatives such as caps/floors and swaptions, however, requires modelling of how the term structure is evolving. More precisely, Bachelier and Black type pricing formulas for caplets/floorlets/swaplets, implicitly assume particular evolutions of the forward rate. The Bachelier type pricing assumes the stochastic differential equation (SDE),

$$dF(t, T_0, T_1) = \sigma dW^{Q^{T_1}}, \tag{3}$$

where σ is a constant volatility and $dW^{Q^T_1}$ is standard Brownian motion under the T_1-forward measure Q^{T_1}. Changes in $F(t, T_0, T_1)$ should therefore follow a normal distribution. Black type pricing assumes

$$dF(t, T_0, T_1) = \sigma F(t, T_0, T_1) dW^{Q^{T_1}}, \tag{4}$$

and changes in $F(t, T_0, T_1)$ should be log normally distributed.

The evolution we observe in the market instead corresponds to the objective forward measure P. It includes effects of risk (e.g. term-premium or even default), grouped under the guise of the market price of risk, and may therefore also result in non-negligible drift terms in the associated SDEs. Thus we will check whether the forward rate evolves according either to a Bachelier type SDE

$$dF(t, T_0, T_1) = \sigma dW^P + \mu dt, \tag{5}$$

or a Black type SDE

$$dF(t, T_0, T_1) = \sigma F(t, T_0, T_1) dW^P + \mu F(t, T_0, T_1) dt, \tag{6}$$

where dW^P is standard Brownian motion under P and μ is the drift. Note the drift term may have time dependence but, for simplicity, we assume constant drift in our initial analysis (it turns out the drift terms will be negligible for our data). To test whether the above SDEs conform to empirical observations, we can use market data, and check whether the statistical properties of the daily changes in $F(t, T_0, T_1)$ support the hypothesis of either Eqs. (5) or (6).

It is often stated in the financial literature and the popular press — and may therefore be considered well-known — that returns on asset prices are fat-tailed and therefore fail to follow the statistics implied by normality or log normality. The above modelling should therefore only be considered a useful starting point. Typically implied volatility surfaces are computed given observed derivative prices, which can in turn be used to interpolate prices for alternative maturities and strikes, or be used in conjunction with the greeks for risk management. Nevertheless, as a useful exercise, we wish to check empirically, that the above SDEs fail to describe market data on forward rates.

2 The Data

The federal reserve provides historical data of $y(t, T)$ observed in the markets, for a range of maturities $T = 1/12, 1/6, 1/4, 1/2, 1, 3, 5, 7, 10, 20, 30$ years.[1] In this analysis we use ten years of data, from $t =$2015-12-22 to $t =$2025-12-22. There are no observations on weekends and holidays. The data allow us to reconstruct the yield curve and discount bond (ZCB) curve as functions of T for any observation date t. We use a simple spline interpolation to compute values that fall between the provided maturities in the data.

Using the ZCB prices, we can find $F(t, T_0, T_1)$ using Eq. (2), for some choice T_0 and T_1. In this analysis we fix, $T_0 =$ 2025-09-22 and $T_1 =$ 2025-12-22, corresponding to a simple forward rate for an approximate three month time frame. We then examine $F(t, T_0, T_1)$ for the period $t =$ 2015-12-22 to $t =$ 2025-09-22. The forward rate is shown in Fig. 1.

[1] See https://fred.stlouisfed.org/release/tables?rid=18&eid=289

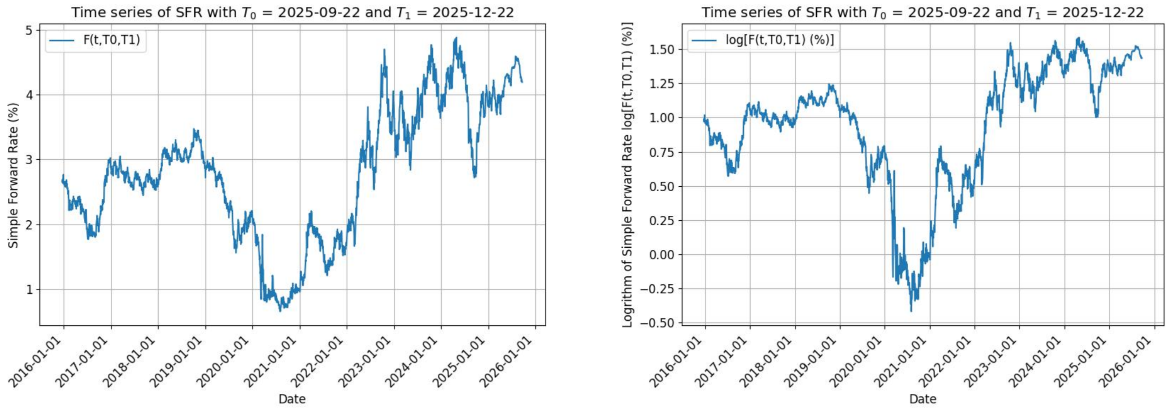

Figure 1: Left: The evolution of $F(t, T_0, T_1)$ with time. Right: The natural logarithm of the same.

Some preliminary observations are in order. Around March 2020 the Federal Reserve aggressively cut interest rates in response to the economic disruption of the Covid pandemic. It further targeted interest rates through its quantitative easing programme. This in turn drove down yields of five year treasury notes and hence we also of our forward rate for borrowing in 2025. Following the pandemic, there was a bout of world wide inflation attributed to supply chain disruptions. The Federal Reserve raised interest rates and also engaged in quantitative tightening (from June 2022), which drove up three year treasury note yields, and hence also the forward rate of interest. We will see the impact of these events reflected in the time dependent volatility calculated in Sec. 4 below.

3 Testing for Normality

For the Bachelier type hypothesis, we examine daily changes of $F(t, T_0, T_1)$, more precisely

$$\Delta F(t, T_0, T_1) \equiv F(t, T_0, T_1) - F(t - 1\,\text{day}, T_0, T_1). \tag{7}$$

Given Eq. (5), one expects a distribution

$$\Delta F(t, T_0, T_1) \sim \mathcal{N}(\mu, \sigma^2), \tag{8}$$

i.e. a normal distribution with daily drift μ and variance σ^2. We will fit the model by finding μ and σ from the data and then testing whether the distribution is indeed Gaussian.

Due to weekends and holidays there are gaps in the data, where multiple calendar days separate actual trading days. A scaling with the difference in calendar days, $\Delta(t)$ would imply an increase in the variance $\sigma^2 \to \Delta(t)\sigma^2$ for such observations. Textbooks such as Hull, however, emphasise that trading itself drives volatility. A priori, we do not know which of the two (indeed if either), correspond to the empirical data. For now, we therefore exclude such observations from our analysis (we will re-introduce them for our volatility forecasting in Sec. Sec:ML). With such dates removed, we have a total of 1903 data points to test the hypothesis.

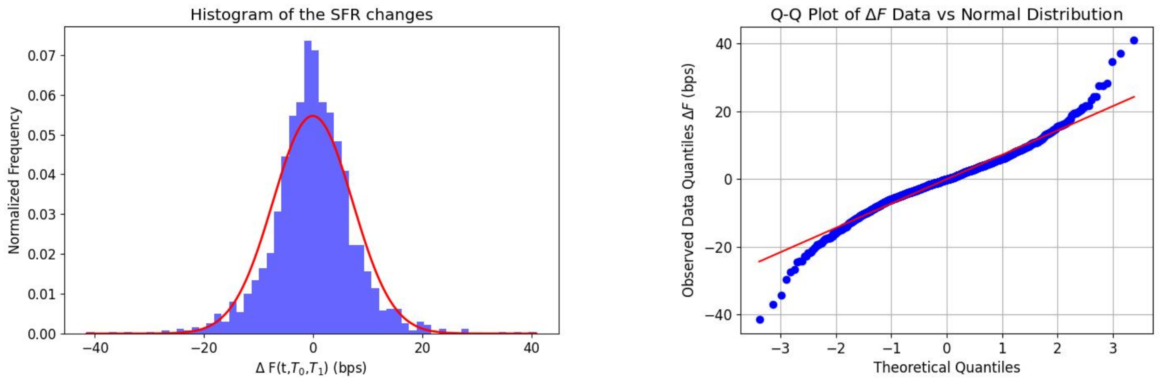

Figure 2: Left: Histogram of $\Delta F(t, T_0, T_1)$ and comparison to the normal pdf. The data looks approximately Gaussian but this is illusionary. Right: Q-Q plot comparing the data with the normal distribution (the zero intercept corresponds to $\hat{\mu}$ and the gradient to $\hat{\sigma}$). Data points on the left of the plot show downward jumps with magnitude much larger than expected from a normal distribution. Similarly data points on the right of the plot show positive jumps much larger than expected from a normal distribution. This is characteristic of a fat tailed distribution.

We now analyse our data. The sample mean and standard deviation for the daily changes ΔF are

$$\hat{\mu} = -0.07 \pm 0.28 \text{ bps} \qquad\qquad \hat{\sigma} = 7.30 \pm 0.19 \text{ bps}, \tag{9}$$

which act as estimates for μ and σ. Errors are quoted to give the 90% confidence interval (CIs). Note the drift term is compatible with zero. A histogram with the observed changes and comparison to a Gaussian pdf fitted with the above μ and σ is shown in Fig. 2. The associated Q-Q plot also shown in Fig. 2. The Q-Q plot indicates fat-tailed departures from normality. This is backed up by various statistical tests for normality (for which we use the scipy.stats and statsmodels.stats python libraries). The Shapiro-Wilk statistic 0.97 corresponds to a $p-$value of $\simeq 10^{-19}$.[2] The Lilliefors KS statistic 0.052 to a $p-$value of $\simeq 10^{-13}$. The Anderson-Darling test statistic of 11 to significance level $\ll 1\%$. All three tests indicate we should reject the null hypothesis of a normal distribution with constant volatility. As a sanity-check, we sort the daily changes in ascending order and find $\Delta F_i = -41.5,\ -36.9,\ -34.3, ..., 34.5,\ 37.0,\ 40.9$ (bps) corresponding to z-values

$$z_i \equiv \frac{\Delta F_i - \hat{\mu}}{\hat{\sigma}} = -5.68, -5.05, -4.69,\ ...\ ,\ 4.74,\ 5.08,\ 5.61. \tag{10}$$

That is, we have four five sigma changes in 1903 data points, although a five fluctuation should only be observed in approximately 1 in 1.7 million data points for a normal distribution. So it makes sense that we reject normality with constant volatility for this dataset.

For the Black type hypothesis, we examine daily changes of $\log F(t, T_0, T_1)$, namely

$$\Delta \log F(t, T_0, T_1) \equiv \log F(t, T_0, T_1) - \log F(t - 1\,\text{day}, T_0, T_1). \tag{11}$$

[2] The Shapiro-Wilk statistic is close to unity but the p-value very small, so we should not rely solely on this test, as Shapiro-Wilk may reject normality for large data-sets even for rather unimportant deviations.

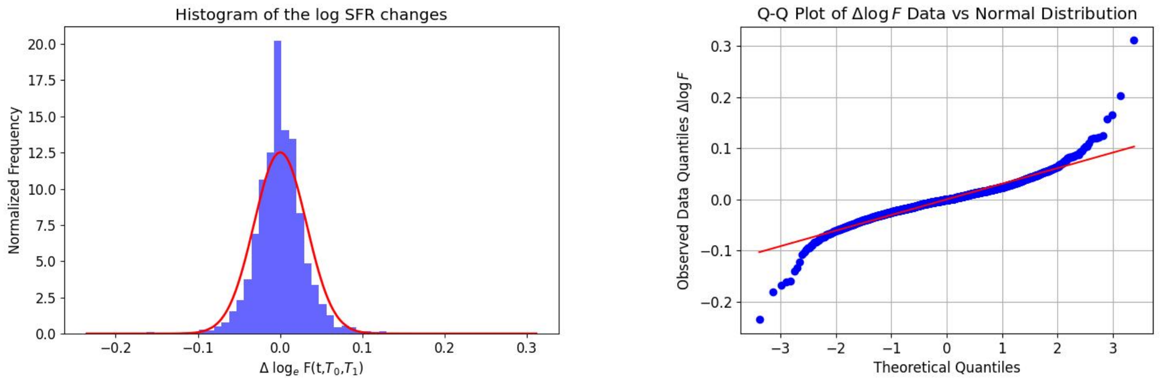

Figure 3: Left: Histogram of $\Delta \log F(t, T_0, T_1)$ and comparison to the normal pdf. Again the data looks approximately Gaussian but this is illusionary. Right: Q-Q plot comparing the data with the normal distribution again showing fat tails.

Given Eq. (6), one expects a distribution

$$\Delta \log F(t, T_0, T_1) \sim \mathcal{N}(\mu - \sigma^2/2, \sigma^2), \tag{12}$$

i.e. a normal distribution with daily drift $\theta \equiv \mu - \sigma^2/2$ and variance σ^2. Again we exclude observations where trading days are separated by more than one calendar day, leaving us 1903 data points. The sample mean and standard deviation are

$$\hat{\theta} = (5.72 \times 10^{-5}) \pm (1.20 \times 10^{-3}), \qquad \hat{\sigma} = (3.18 \times 10^{-2}) \pm (8.50 \times 10^{-4}). \tag{13}$$

The drift term is again compatible with zero. The pdf and Q-Q plot for the Black type hypothesis is shown in Fig. 3, showing a fat-tailed distribution, departing significantly from normality. Here, the Shapiro-Wilk statistic 0.92 corresponds to a $p-$value of $\simeq 10^{-30}$. The Lilliefors KS statistic 0.070 to a $p-$value of $\simeq 10^{-24}$. The Anderson-Darling test statistic of 22 to significance level $\ll 1\%$. All three tests indicate we should reject the null hypothesis of a normal distribution with constant volatility. The sorted jumps in ascending order are $\Delta \log F_i = -0.24,\ -0.18,\ -0.17, ...,\ 0.17,\ 0.20,\ 0.31$. The associated z-values are

$$z_i \equiv \frac{\Delta \log F_i - \hat{\theta}}{\hat{\sigma}} = -7.37, -5.67, -5.25,\ \dots,\ 5.19,\ 6.35,\ 9.75. \tag{14}$$

Here we again observe far larger jumps than expected for a normal distribution, including a 9σ fluctuation, which would only appear once in $\sim 10^{-19}$ observations for a normal distribution.

To check that nothing funny is happening in our interpolation of the ZCB curves — leading to spurious jumps in the data — we plot the ZCB curves corresponding to the largest jumps and perform a visual check. Example plots, showing the spline interpolation is working as expected, are shown in Fig. 4.

Next we check whether the first half of the dataset and the second half of the dataset are statistically likely to be coming from the same distribution (independently of any theoretical idea as to whatever that distribution may be). We perform Two-Sample Kolmogorov-Smirnov (KS) Tests for the two halves of the Bachelier (ΔF) and Black type ($\Delta \log F$)

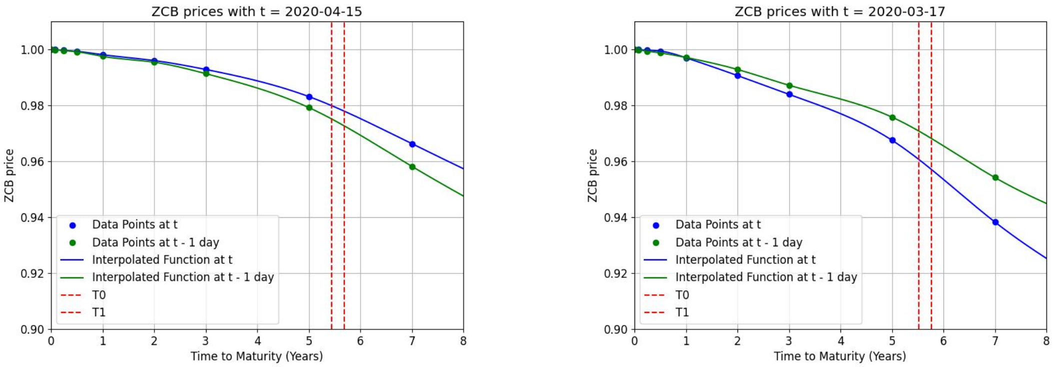

Figure 4: Left: ZCB (discount) bond curves corresponding to the largest negative $\Delta \log F(t, T_0, T_1)$ showing the behaviour of the spline interpolation. The vertical lines correspond to our choice of reset and maturity date T_0 and T_1 for the forward rate. Right: ZCB (discount) bond curves corresponding to the largest positive $\Delta \log F(t, T_0, T_1)$.

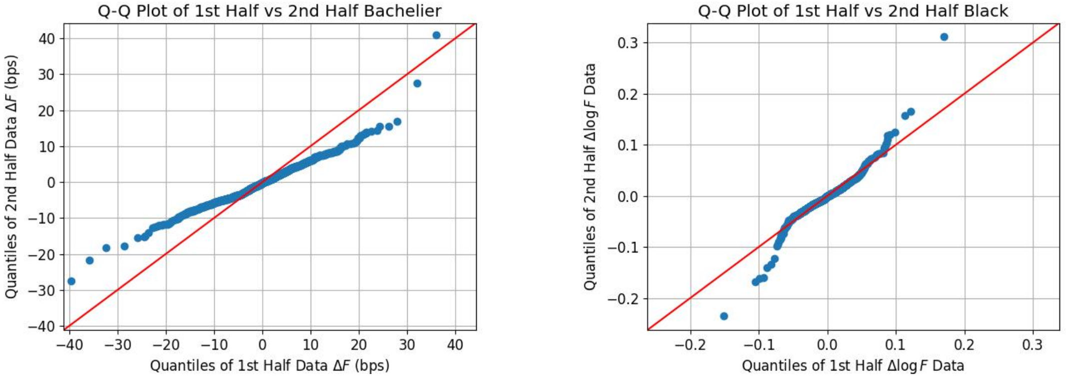

Figure 5: Left: Two sample Q-Q plot for $\Delta F(t, T_0, T_1)$ after splitting the data into two halves. Right: Two sample Q-Q plot for $\Delta \log F(t, T_0, T_1)$ after splitting the data into two halves. Significant deviations of the points from the diagonal indicate the samples are not drawn from the same distributions.

datasets. For the Bachelier data we obtain a KS statistic of 0.11, corresponding to $p \simeq 10^{-5}$, rejecting the null hypothesis that the first and second half of the data can be described as coming from the same distribution. For the Black data we obtain a KS statistic of 0.076 and corresponding $p \simeq 6.8 \times 10^{-3}$, also rejecting the null hypothesis. We also compare the two halves in Q-Q plots shown in Fig. 5, visually showing the first and second halves differ.

Finally, we examine the scaling in the volatility as the time period of observation increases. For full trading weeks, we extract weekly changes in $F(t, T_0, T_1)$ and $\log F(t, T_0, T_1)$, i.e. Friday minus Monday values ($\Delta t = 4$ days). Each of the dataset has 402 observations. We find the $\Delta t = 4$ day drift remains consistent with zero. For Brownian motion, we expect the $\Delta t = 4$ volatility to be a factor of two larger than the $\Delta t = 1$ day volatility we found earlier. As a naive metric, we compare the 90% CIs of the weekly volatility with the 90%

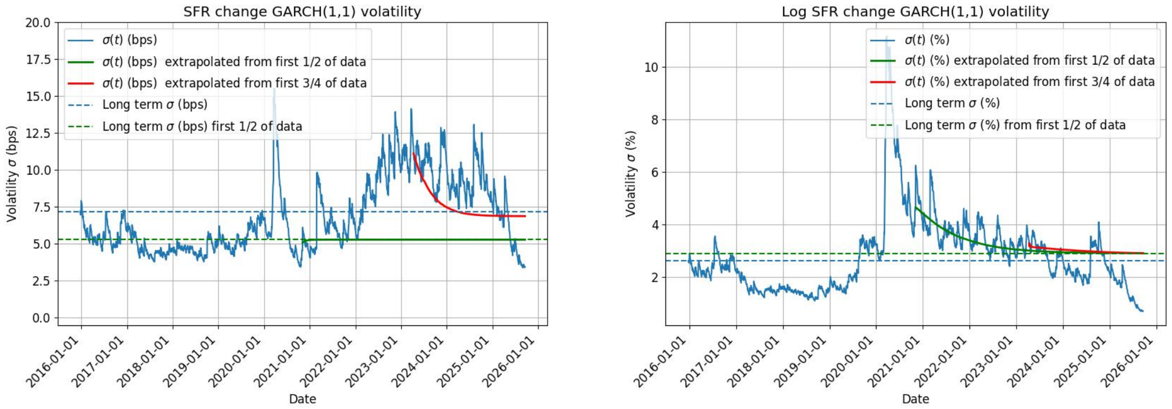

Figure 6: Left: GARCH(1,1) analysis of the $\Delta F(t, T_0, T_1)$ and forecasts using subsamples of the data. We observe the rapid increase in the volatility around the start of the Covid pandemic and also in 2022 when rates rose in response to inflation. The first half GARCH forecast returns to a lower estimate of the long term volatility. Right: Similar but for $\Delta \log F(t, T_0, T_1)$. We see the spike for Covid but less of a sustained increase in the volatility after 2022 because taking the logarithm makes us less sensitive to jumps in raw bps when rates are already high. Both GARCH forecasts do a reasonable job visually for the logarithmic case.

CI's of the daily volatility multiplied by two, both for the Bachelier

$$13.01 < \sigma_{4\text{day}} \text{ (bps)} < 14.61, \qquad 14.21 < 2 \times \sigma_{1\text{day}} \text{ (bps)} < 14.99, \tag{15}$$

and the Black type hypothesis

$$0.061 < \sigma_{4\text{day}} < 0.069, \qquad 0.062 < 2 \times \sigma_{1\text{day}} < 0.065. \tag{16}$$

Interesting, we do indeed find an overlap in both cases. Note the range is much larger or the direct 4-day estimate, because of the smaller sample size (402 vs 1903 observations).

4 GARCH(1,1) Analysis for the Volatility

Time dependence in the volatility and volatility clustering is a common phenomenon in financial data. To model the volatility, GARCH analyses are commonly used. Here we use a GARCH(1,1) analysis where the volatility on trading day i depends on a long term average volatility, the volatility on trading day $i-1$, and the change in the variable of interest at $i-1$, u_{i-1}, here taken to be either ΔF_{i-1} or $\Delta \log F_{i-1}$, where we multiply the latter by 100 to recover a "percentage" change.[3]

[3] The percentage here corresponds to the relation $F_{i-1} = F_{i-2} \times \text{Exp}([u_{i-1} \text{in}\%]/100)$. Often in the literature, the percentage fractional change is used, $u_{i-1} \equiv 100 \times (S_{i-1} - S_{i-2})/S_{i-2}$, for a GARCH analysis of underlying asset S. We have checked doing so for the forward rate is qualitatively similar to our analysis with $\Delta \log F_{i-1}$. Here, we stick with using ΔF_{i-1} and $\Delta \log F_{i-1}$, in order to provide continuity with the SDEs in Sec. 1 and the analysis in Sec. 3.

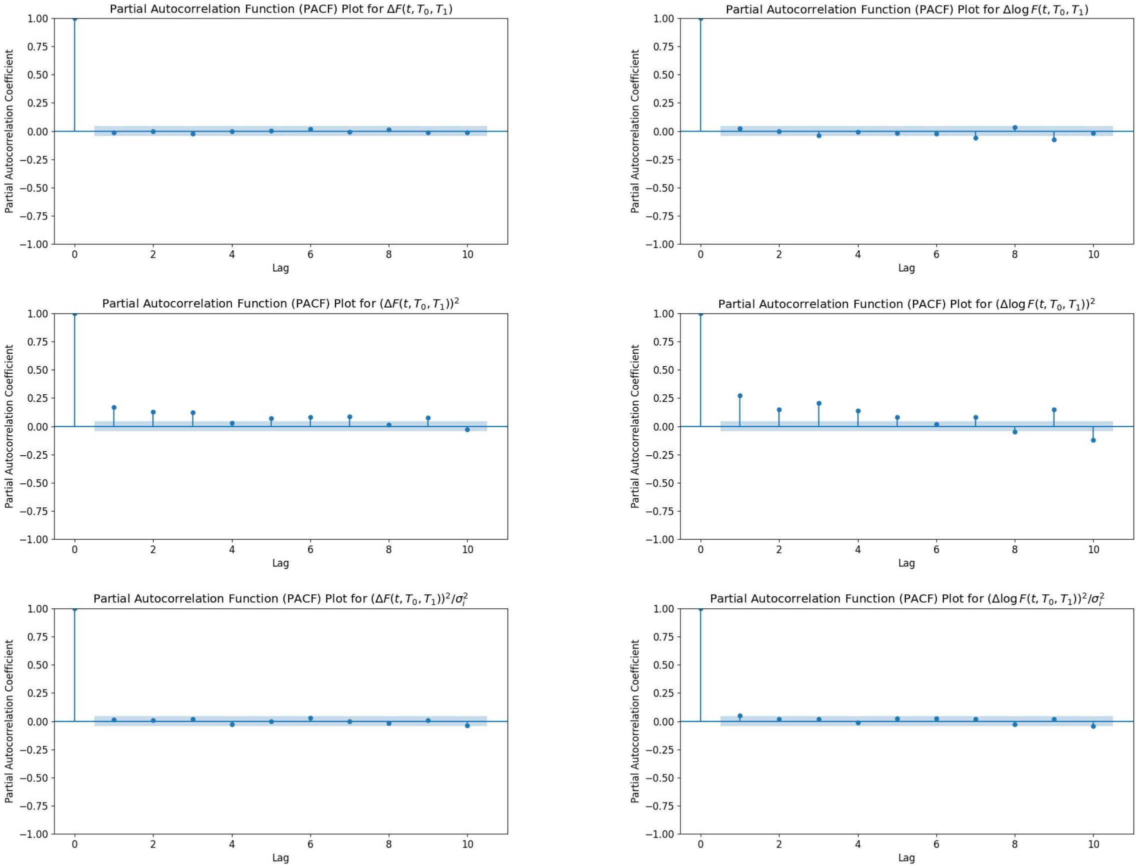

Figure 7: Left: Parital autocorrelations for ΔF_i, $(\Delta F_i)^2$, and $(\Delta F_i)^2/\sigma_i^2$ time series. Right: $\Delta \log F_i$, $(\Delta \log F_i)^2$, and $(\Delta \log F_i)^2/\sigma_i^2$. The squared values display some autocorrelations which, as expected, are effectively removed by dividing by the implied volatility squared.

Mathematically the squared volatility on day i is then

$$\sigma_i^2 = \omega + \alpha u_{i-1}^2 + \beta \sigma_{i-1}^2, \tag{17}$$

where $V_L = \omega/\gamma$ is the long term averaged volatility squared, $\gamma = 1 - \alpha - \beta$, and the ω, α, β are extracted from the data using a maximum log-likelihood. More precisely, the function we wish to maximize is

$$\log \mathcal{L}(\alpha, \beta, \omega) = -\sum_{i=1}^{m} \left[\log(\sigma_i^2) + \frac{u_i^2}{\sigma_i^2} \right], \tag{18}$$

where $i = 1, ..., m$ is the index of our observations over the trading days. We use the standard scipy optimization library to find α, β, ω. We also checked our result agrees with that obtained from the arch_model class from the arch library.

It is also possible to forecast the volatility using GARCH. The forecast volatility follows an exponential decay back to the long term average, with a data-driven rate of the decay (namely dependent on the best-fit α and β). The forecast volatility for day $i + j$, given the

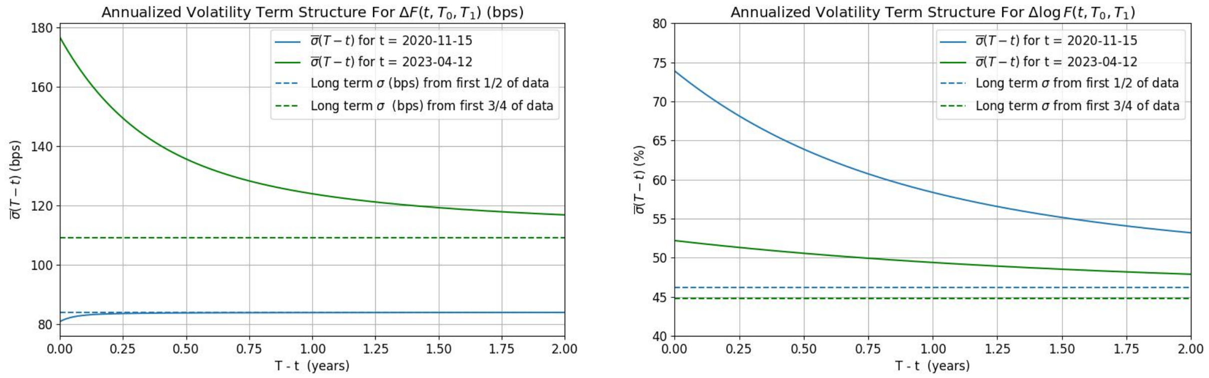

Figure 8: Estimated annualized volatility term structure for ΔF_i and $\Delta \log F_i$, using observations at days 1/2 and 3/4's of the way through the dataset.

volatility on day i, is given by

$$\mathbb{E}[\sigma_{i+j}^2] = V_L + (\alpha + \beta)^j(\sigma_i^2 - V_L) \tag{19}$$

In Fig. 6 we show our calculation of the volatility for both $\Delta F(t, T_0, T_1)$ and $\Delta \log F(t, T_0, T_1)$ along with GARCH predictions extrapolated from either the first 1/2 or 3/4's of the dataset (to do this in a more realistic manner, we do a separate GARCH for each of the smaller datasets, and extrapolate from their final value using their own α, β, ω).

The time series exhibit an interesting non-linear autocorrelation: the values ΔF_i have only a negligible autocorrelation, but the time series of the values squared $(\Delta F_i)^2$ are autocorrelated. In other words, when the volatility jumps, the up and down movements of ΔF_i tend to cancel out, but if $(\Delta F_i)^2$ is elevated then the points around it tend to be also. Similarly for $\Delta \log F_i$ and $(\Delta \log F_i)^2$. This is precisely the volatility clustering visible in our GARCH analysis. To visualise this behaviour, we show partial autocorrelation plots for the time series, the time series of values squared, and the time series of values squared divided by the GARCH volatility squared in Fig. 7 (using the statsmodels library). Notice the effect of the GARCH volatility is to fit the required σ_i^2 to remove the autocorrelations in the latter time series.

The estimated annualized volatility term structure, assuming 252 trading days per calendar year, is given by (e.g. see Hull 11th edition, Eq. [23.14]),

$$\sigma(T) = \sqrt{252\left(V_L + \frac{1 - e^{-aT}}{aT}[\sigma_i^2 - V_L]\right)} \tag{20}$$

where $a = -\log(\alpha + \beta)$, σ_i is the current volatility, and T is the maturity from the current day. It is an estimate of the average volatility to the maturity using the GARCH forecast. Plots of the volatility term structure are shown in Fig. 8. According to Hull, the estimated volatility term structure can be used to prognosticate how changes to the current volatility will impact the implied volatility surface. (The implied volatility comes from observations of market prices for derivatives, assuming a given pricing model, rather than time series anaylsis of the underlying asset price).

5 Machine Learning for Volatility Forecasting

We have now seen how GARCH can be used to calculate the volatility and also provides a simple forecast. We now ask whether machine learning techniques can be useful in offering alternative volatility forecasts. For this section, we also include trading days after weekends and holidays, to make our forecasts more applicable to actual trading. We also limit ourselves to the volatility of $\Delta \log F_i$ for brevity.

To avoid leakage between the training, validation, and test datasets we first compute a running GARCH(1,1) volatility using only the data up to the present day for the optimization. Except for the first 1000 days of data, for which we use all the data up to day 1000, as we required a certain number of data points to do the GARCH analysis (there is no leakage issue here because all these days will be in our training set). One could imagine that the running GARCH would be updated daily as new data trading data arrives, beyond our eventual analysis, and so it also better mimics real world applications.

Having defined our volatility timeseries, we now seek to train neural networks to forecast the volatility into the future. We will be ambitious, and seek to forecast ~ 3 calendar months into the future, i.e. 63 trading days. As an input, we will give the volatility of the previous 63 trading days. We then train two neural networks. First a neural network which passes the input through one or more dense layers and as output gives an estimate of the volatility for each of the next 63 days. We label this one NN. Secondly, a network using Long-Short-Term-Memory with an attention mechanism and encoder/decoder sequence to sequence structure (which we label LSTM).

NN Hyperparameter	**Search Range/Options**	**Best Value**
Number of Layers	1–3	3
Units per Layer	64–320 (Step: 64)	256, 128, 256
Batch Normalization	True/False (Per layer)	F/F/T
Dropout	True/False (Per layer)	T/T/F
Dropout Rate	0.0–0.5 (Step: 0.1)	0.3/0.3/NA
Learning Rate	10^{-5}–10^{-3} (Log sampling)	4.26×10^{-5}
Batch Size	8–32 (Step: 8)	24

Table 1: NN Hyperparameter search space and optimal values found. The optimal values give a network with ~ 99000 *tunable parameters.*

The data is split 60/20/20 into training, validation, and test sets. Buffers are created around each to prevent overlap of the lags and future volatility data of each. Similarly there is no shuffling of the data between training, validation, and test sets as this would cause leakage (note the sequence of volatilities on day i, with its lags and future features, of course looks very similar to the sequence of volatilities on day $i+1$). For the LSTM we also include a smoothed Exponentially Weighted Moving Average (EWMA), with decay parameter $\lambda =$

LSTM Hyperparameter	**Search Range**	**Best Value**
Number of Encoding Layers	1–3	1
Units	32–256 (Step: 32)	32
Dropout	0.1–0.4 (Step: 0.1)	0.1
Learning Rate	5×10^{-6} – 5×10^{-4} (log scale)	3.24×10^{-4}
Batch Size	16, 32, 64, 128	16

Table 2: LSTM Hyperparameter search space and optimal values found. The optimal values give a network with ~ 13000 *tunable parameters.*

0.99, over the lags as an additional feature. After datasets are split, buffered, and transformed into the correct format for the NN and LSTM, we are left with training/validation/test set lengths of 1260/399/400 for the NN and 1336/363/363 for the LSTM. Before input into the model, the NN data is scaled using the StandardScaler from the sklearn library, and the LSTM data is scaled using the RobustScaler from the same library.

We use the python libraries tensorflow, keras, and keras_tuner, to construct, train, and scan over hyper-parameters and architectures (learning rate, batch sizes, batch normalization, dropouts, layer numbers, neurons per layer) to find the best choices that minimizes the loss over the validation set. Some experimentation showed improved behaviour when minimizing over the mean absolute error rather than the mean squared error. A hypothesis for this is that the large spike in volatility present in the training data around the Covid pandemic, has an overly large impact on the model training when squared errors are used.

The range of hyper-parameters we search over and optimal values are shown in Tables 1 and 2. In our search we used 150 trials for the NN and 52 trials for the LSTM (the latter requires more training time per trial). When searching, we allow for 500 epochs of training for each hyperparameter set, although in practice this is never reached, due to a callback mechanism for the validation loss with a patience of 100 (50) for the NN (LSTM).

After optimizing over the hyper-parameters, we train the networks a final time using the best hyper-parameters (increasing the patience to 100 for the LSTM). Training and validation loss curves are shown in Fig. 9. We also compare the MAE and MSE of the predictions over the test set and contrast with the MAE and MSE of the GARCH forecast in Table 3. Plots showing the forecasts for some choices of sequences in the test set are shown in Fig. 10. We see the NN provides a rather noisy forecast, with an overall trend somewhat more suppressed than the GARCH benchmark. The LSTM on the other hand seems to have found a solution with an initial jump like structure followed by a long plateau. Quantitatively, it does somewhat better than the GARCH and NN predictions over the test data.

We see that GARCH continues to perform well over this dataset, but with key qualitative differences between the three types of predictions. Note the number of sequences with no overlaps in the training set is rather small $\sim 1300/126 \sim 10$, so there is no large surprise that the current set-up fails to perform significantly better than GARCH. Realistically, we require larger datasets for such inherently noisy financial data, which would also allow for additional improvements such as error estimates to be included (via conformal prediction). This would perhaps also allow alternative estimates of the volatility term structure, using

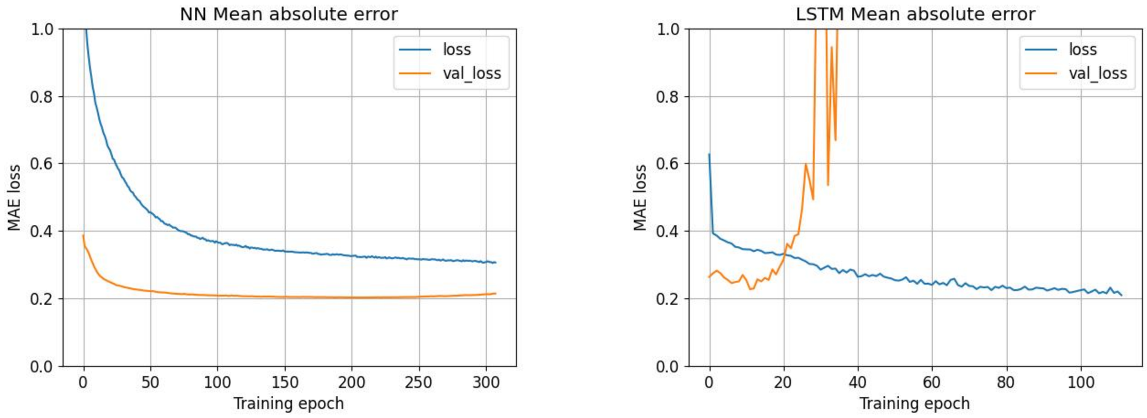

Figure 9: Training and validation loss curves using the scaled datasets with the optimal hyperparameters for the NN (left) and LSTM (right) as a function of the training epoch. As we can see, for the optimum hyper-parameters found in our search algorithm, the LSTM training is rather noisy. Note the validation loss may be consistently smaller than the training loss due to the latter being calculated with dropout and averaged over the batches during the training epoch.

Model	**MSE**	**MAE**
GARCH (NN Test Data)	0.33	0.45
NN	0.37	0.46

Model	**MSE**	**MAE**
GARCH (LSTM Test Data)	0.35	0.47
LSTM	0.32	0.43

Table 3: Model Performance Comparison on Test Data. The Test data is of slightly different lengths for the two networks, after the data sets are split and transformed into the correct format, so we have differing GARCH errors.

the machine learning predictions over a longer time frame, rather than GARCH.

6 Conclusions

We demonstrated empirically that the simple forward rate cannot be described using a constant volatility, as assumed in common pricing benchmarks (although these pricing formulas remain useful). We used a GARCH analysis to estimate and forecast the volatility. We discussed how the estimated volatility term structure can help calibrate movements in the implied volatility curve (derived from derivative market prices). We then used machine learning in an exploratory effort for providing alternative volatility predictions. With additional training and testing, such techniques may prove useful alternative forecasts and error estimates, and for understanding and calibrating movements of the implied volatility surface (for trading and risk management purposes).

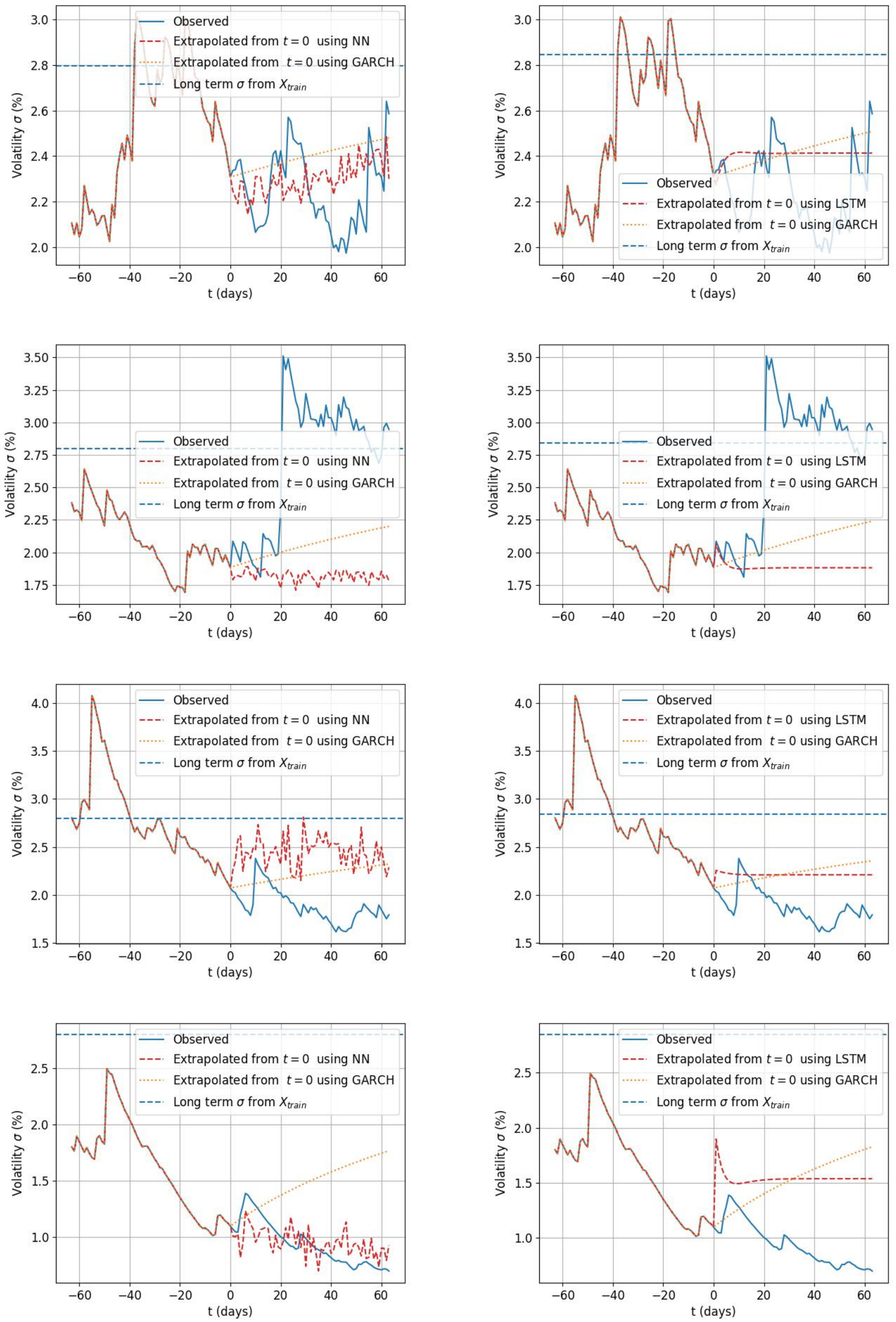

Figure 10: Some example inputs and predictions of the volatility from the test data. Predictions on the left (right) column are from the NN (LSTM). The dates are chosen so that the example on row i + 1 is 120 trading days advanced from example i. The last example corresponds to the very end of the observations of the timeseries.