# Predicting volatility using the term structure curves of commodity futures

Master's Thesis Hidde Hogenhout

Hidde Hogenhout (500865hh)

Supervisor: Onno Kleen

Second Assessor: PA Opschoor

Company supervisors: Michael van Enkhuizen & Mattijs Oort

March 24, 2024

## Abstract

We improve volatility predictions for 19 commodity markets by adding the shape of predicted term structure curves to a standard volatility model. The term structure curves are predicted through a Nelson-Siegel style model that has been adapted with a seasonal component. A parsimonious version of the model called daily Nelson-Siegel provides optimal fit to the price curve on a single day. It is compared to a dynamic Nelson-Siegel version that is estimated over a longer period, allowing it to also model the changes from day to day. The daily model produces more accurate predictions, but is prone to overfitting to the observed prices rather than correctly predicting the shape of the whole curve. It is shown that this leads to unstable and very large estimates of the hidden shape factors. Therefore, shape factors predicted using the state-space methodology are generally more useful in augmenting the baseline heterogeneous autoregressive (HAR) volatility model. Some proposed model specifications significantly beat this baseline for various commodities, showing that there is potential for the technique. Furthermore, we find that a LASSO regularised regression based on all predicted hidden shape factors can significantly improve volatility forecasting in the natural gas market.

ERASMUS UNIVERSITEIT ROTTERDAM

transtrend

## Acknowledgements

While this thesis is the culmination of my studies and my final work, I would never have been able to do it all alone. I would like to thank Transtrend, where I have felt welcomed since day one, for giving me the opportunity to do this research. Specifically, my supervisors Michael and Mattijs for teaching me about futures markets, for having fate in me when I needed it, and for the good times outside of research. In addition, I would like to express my gratitude to Onno Kleen for consistently finding time to meet with me, despite his recent transition to fatherhood and likely busier schedule than mine. Thank you to my parents, brother, and other family for always being there for me, providing a safe haven which I can always count on, no matter what happens. Finally, I would like to thank my friends for making sure I know there is more to studying then doing research.

# Contents

# 1 Introduction

Every product we use or item of food we eat is made in some way from raw commodities. Therefore, the prices of these commodities greatly influence the prices of all the goods we consume. Being able to predict their dynamics could help prepare the world for sudden price increases, limit risk exposure in a transition to clean energy, and even detect pending food shortages early. This article aims to predict volatility in commodity markets using the shape of their price curve as a predictive signal. The price curve, also named the term structure in this paper, refers to how prices differ for futures contracts on the same commodity with varying times-to-maturity, much like the yield curve for bonds. The structure of this curve is determined by supply and demand factors of the physical markets underlying the futures contracts. For example, a supply shortage which is expected to last for 6 months will have the effect of increased prices for all futures contracts that expire within those months, leading to a downward-sloping term structure curve. In addition, shocks to supply and demand often come paired with increased volatility in that market. Therefore, we expect that if the shape of the curve can be predicted, it will contain information about volatility in the underlying market.

In order to use future term structure curves to predict volatility in commodity markets, these curves first have to be predicted. To do so, a Nelson-Siegel (Nelson & Siegel, 1987) style model is implemented on the price curves of the futures. The model is augmented with a seasonality component to better fit the dynamics of commodity markets, such as harvesting periods and winter heating demand. Furthermore, it is estimated both through an approach that maximises the fit in each individual day and using state-space methodology to also model the changes from day to day. The methods are tested in an extensive out-of-sample (OOS) prediction exercise that will provide an answer to the first research question: *To what extent is it possible to predict the shape of a commodity futures price curve using only past price information?*

In the second part of this paper, these price curves are used to predict volatility. A standard heterogeneous autoregressive (HAR) (Corsi, 2009)volatility model is used as a baseline and subsequently improved using aspects of the predicted curves, such as their slope and curvature. Furthermore, some basic machine learning (ML) techniques are used to investigate possible non-linear relationships between the shape of the curve and volatility. All volatility models are compared in an out-of-sample prediction exercise. In this section of the research, the central question is: *To what degree can predictions by a standard volatility model be improved by using the predicted shape of the term structure in commodity markets?*

The academic literature on commodity futures and their term structure curves dates back to Keynes (1923), who, together with Hicks (1939), came up with the theory of normal back-

wardation. They argue that producers of a commodity, such as farmers producing wheat, want to hedge their price risk and are willing to pay a risk premium to do so. The longer the time-to-maturity of a contract, the higher the risk, so the premium paid will be larger. Since producers are sellers of the commodity, the premium they pay will decrease the price of the commodity, leading to decreasing prices as time-to-maturity increases. This is known in commodity markets as backwardation. However, the theory of normal backwardation was not able to (and never meant to) explain the fact that other commodity markets often display a contango term structure, characterised by increasing prices as time-to-maturity increases. To account for this phenomenon, Hirshleifer (1989, 1990) extends the theory of normal backwardation by stating that the risk premium could also be paid by the users of the commodity wishing to hedge the price risk in their supply. His generalised hedging pressure theory states that whichever group, producers or users, has the largest interest in hedging risk, will be paying the risk premium. The risk-premium can then thus be positive or negative, leading, respectively, to increasing or decreasing prices as time-maturity increases. Stated otherwise: If the producers of a commodity are more inclined to hedge, the term-structure will be backwardation, while it will be in contango if the users have larger incentive to hedge.

Kaldor (1939) produced an alternative theory to explain the term structures of commodity markets, which was further developed by Working (1949) and Brennan (1976). The storage costs theory argues that storing a commodity is often not free, for example, because warehouses require rent to be paid, and agricultural commodities have limited lifespans. This cost of storing the commodity will be reflected in a higher price for longer maturities, thus explaining contango term structures. The theory of storage deals with markets in backwardation by arguing that there is also a benefit to holding a commodity, defined as the convenience yield. The convenience yield can be explained as the value of having the real option to sell the commodity at any time while it is in your storage, comparable to the premium on financial options contracts. In the event of a sudden negative shock to supply or a positive shock to demand, the holder of the commodity will be able to sell at a premium and thus make a profit. This leads Deaton and Laroque (1992) to formulate the hypothesis that the convenience yield is negatively related to inventory levels, since low levels of inventory increase the probability of a supply shortage. They further developed the idea in Deaton and Laroque (1996) and it was finally theoretically proven by Routledge et al. (2000). Many studies, such as Gorton et al. (2013) and Pindyck (1990), showed the relationship between inventory levels and convenience yield empirically.

The practice of modelling and predicting term structure curves arose from a desire to capture the entire set of bond yields using only a few parameters. Nelson and Siegel (1987) is a

revolutionary paper in this space, describing how the yield curve can be modelled using only three estimated coefficients on regressors that depend on maturity. The model is further developed in Diebold and Li (2006), who also propose the alternative interpretation of the estimated coefficients as time-varying level, slope and curvature factors, respectively. This model will serve as the basis for all models used in this paper. The next big advance comes from Diebold et al. (2006) who propose a state-space formulation of the model with the assumption of the level, slope, and curvature factors having autoregressive properties. Rather than having to estimate each day separately using the cross-section of bond yields, this approach allows for a single model to be estimated using the entire time-series and cross-sectional dimension.

The application of the Nelson-Siegel (NS) model to commodity futures prices is a relatively novel area of the literature. The first to implement this idea is West (2012), who applies the original NS model to cotton, corn, and sugar futures, with the aim of accurately predicting the prices of over-the-counter forward contracts with maturities longer than their exchange-traded counterparts. The paper also adds a seasonal component to the model to deal with the inherent seasonality that arises from the harvesting of agricultural commodities. However, by building on the original Nelson and Siegel (1987) version of the model rather than the Diebold and Li (2006) version, the model loses much of the interpretability that comes with the latter specification. The first to use the Diebold and Li (2006) specification to model the term structure of commodity futures is Heidorn et al. (2015), who use the better interpretation of this model to analyse whether the increased influence of financial firms in the oil market increased the general level of prices. Grønborg and Lunde (2016) apply the same model to the same market, but explicitly evaluate its ability to produce accurate forecasts for the price curve. Karstanje et al. (2017) apply the NS methodology to a large set of commodities to investigate joint effects. Furthermore, both introduce a way to account for seasonality in those markets and apply the state-space methodology by Diebold et al. (2006) to estimate the model for their entire dataset at once. Finally, Bianchi et al. (2023) use the assumption that the change in factors from $t$ to $t+1$ is equal to the change from $t-1$ to $t$, to develop a profitable trading strategy based on the slope factor.

To model volatility, we will rely on the parsimonious yet accurate heterogeneous autoregressive (HAR) model proposed by Corsi (2009). In its general form, it is an AR-type model of realised volatility over different horizons. However, due to data limitations, this paper implements a form based on the true range of the observed prices, rather than the realised volatility. The HAR model is then augmented with various possible predictors based on the shape of the estimated term structure curves. Each of these predictors is used individually, as well as an

implementation containing all of them. The model containing all predictors is estimated not only through standard OLS, but also using some standard implementations of machine learning (ML) techniques. The LASSO estimation by Tibshirani (1996) allows the model to shrink the coefficients of certain predictors to zero, reducing the noise created by introducing all predictors into a single model. Furthermore, the neural network and random forest are able to detect complex and nonlinear relations between the shape of the curve and volatility.

Both the daily and state-space estimation approaches are found to have merits. The daily model produces the most accurate one-day-ahead price predictions for all commodities and detects reasonable seasonal effects. This can be used in practice to investigate whether seasonal effects change over time, for example, to see if climate change has influenced the harvesting moments of various agricultural commodities. However, the daily model suffers from overfitting to the prices observed on the day it is estimated, sometimes leading to correct price predictions but extremely unreasonable estimates for those parts of the curve that are not observed. On the other hand, the state-space model produces less accurate price predictions and is not always able to correctly identify the seasonal effect, but in general predicts a more reasonable and robust shape for the entire curve.

The added robustness of the state-space estimation method proves useful when trying to predict the volatility OOS. The standard volatility model used is unable to handle the extreme shape estimates that are sparsely produced by the daily model, leading to infrequent but very large prediction errors. The state-space estimation method does not suffer from these extreme shape estimates and therefore the volatility model based on this estimation method produces better predictions. Yet, it is noted that the standard volatility model is hard to beat and the standard models can only do so for a few commodities. ML methods are not much better, except for the LASSO regression, which achieves a 7% reduction in mean absolute error compared to the baseline model on the one-month prediction horizon for natural gas.

Although the results presented in this paper are not decisive, they can be used as the basis for a broad range of future commodity futures research. An academically novel way to select actively traded contracts is introduced based on their actual observed activity. Furthermore, we show that with some slight changes there is potential in modelling the term structure of commodity futures using Nelson-Siegel-type models. These models can contemporaneously describe the curve using only a few parameters, rather than all futures prices. This achieves both a dimensionality reduction and adds interpretability to the term structure curve. Furthermore, we show that there are some periods where the price curve forecasting is accurate, showing the potential of these techniques in forecasting price movements. Lastly, we show that the shape

of the curve can be used to improve forecasts of market volatility using a very basic volatility model, paving the way for further research using more advanced volatility models and estimation methods.

The rest of this paper is structured as follows. Section 2 discusses the dataset used and the novel method of cleaning the data, moreover, some example term structure curves are shown. Section 3 describes all models and evaluation metrics used throughout the paper. Section 4 explains the results both quantitatively and graphically. Finally, Section 5 looks back on the main conclusions and provides possible improvements as well as suggestions for further research.

## 2 Data

The data used in this study consist of daily commodity futures contracts' high, low, and closing prices, open interests, trading volumes, and their last trading dates. We study nineteen commodities, which can be divided into five groups and are all traded on US exchanges. Not all data are available from the same starting date, mostly because open interests and trading volumes only started being recorded later. However, the final dataset is one of the most extensive in the current academic literature for commodity futures, featuring over 50 years of daily observations for most commodities. All data used in this research have been compiled from Refinitiv.

Table 1 shows an overview of all commodities and their relevant properties. Most agricultural commodities are available from 1970 to 2023, giving more than 50 years of observations. Natural gas has the shortest sample, only becoming available in 1990. However, 30 years of price data is still more than sufficient. Table 1 also reports the minimum and maximum amount of active contracts on a day, which contracts are defined as active is explained in detail in Section 2.1. Additionally, Figure 2 in Appendix B plots the number of active contracts over time for each commodity.

**Table 1. Summary statistics for all commodities in the sample**

| | | | | | | Actives | | | |
|---|---|---|---|---|---|---|---|---|---|
| Group | Commodity | Short | Start | End | T | Min | Max | Months | Exchange |
| Agries | Chicago Wheat | W | 1970-01 | 2023-06 | 13475 | 3 | 9 | 5 | CME |
| | Kansas Wheat | KW | 1980-01 | 2022-06 | 10712 | 3 | 9 | 5 | CME |
| | Corn | C | 1970-01 | 2023-06 | 13475 | 4 | 10 | 5 | CME |
| | Soybeans | S | 1970-01 | 2023-06 | 13473 | 4 | 11 | 7 | CME |
| | Soybean Meal | SM | 1970-01 | 2023-06 | 13473 | 5 | 13 | 8 | CME |
| | Soybean Oil | BO | 1970-01 | 2023-06 | 13473 | 5 | 13 | 8 | CME |
| Meats | Feeder Cattle | FC | 1981-01 | 2023-06 | 10717 | 4 | 8 | 8 | CME |
| | Live Cattle | LC | 1981-01 | 2023-06 | 10716 | 4 | 8 | 6 | CME |
| | Lean Hogs | LH | 1981-01 | 2023-06 | 10716 | 4 | 10 | 8 | CME |
| Energy | Crude Oil | CL | 1983-03 | 2023-06 | 10136 | 4 | 32 | 12 | CME |
| | RBOB Gasoline | RB | 1984-12 | 2023-06 | 9713 | 3 | 18 | 12 | CME |
| | Natural Gas | NG | 1990-04 | 2022-06 | 8101 | 7 | 57 | 12 | CME |
| | Heating Oil | HO | 1980-01 | 2023-06 | 10950 | 3 | 27 | 12 | CME |
| Metals | Copper | HG | 1973-09 | 2023-06 | 12544 | 3 | 20 | 12 | CME |
| | Gold | GC | 1975-01 | 2023-06 | 12216 | 2 | 17 | 8 | CME |
| Softs | Arabica Coffee | KC | 1975-01 | 2023-06 | 12152 | 3 | 14 | 5 | ICE |
| | Cotton | CT | 1970-01 | 2023-06 | 13422 | 3 | 10 | 5 | ICE |
| | Sugar | SU | 1980-01 | 2022-06 | 10662 | 3 | 12 | 4 | ICE |
| | Cocoa | CO | 1970-01 | 2023-06 | 13397 | 3 | 10 | 5 | ICE |

*Note.* Table reports summary statistics for all commodity futures in the dataset, grouped by their sectors. Short denotes the mnemonic used for the commodity throughout the paper. Start and end indicate the start and end of the dataset respectively, while T gives the total number of time-series observations. Min and max denote the minimum and maximum number of active contracts on a day, months indicates how many contracts are listed on the exchange within a year. Finally, exchange gives the exchange which the commodity currently trades on, note that this could have historically been a different exchange.

## 2.1 Data Cleaning

On any given date, the exchange publishes closing prices for a large number of contracts. For the most popular commodities with a contract each month, such as natural gas or crude oil, there are more than 100 contracts with a listed closing price each day. However, it is important to realise that most of the trading happens in only a small subset of these contracts. For all other contracts, the exchange ensures that the published price falls within some reasonable range based on the prices of the actively traded contracts. To ensure that the prices used in this paper are determined by market forces alone rather than some exchange imposed model, we will impose a lower limit on the traded daily volume for the contract to be considered active. The daily lower volume limit for a commodity is set at 1% of the traded volume of the most actively traded futures contract on that commodity. A contract that achieves this limit once will be considered active for all following days, even if it might drop below the lower limit afterward, ensuring a continuous price series for all contracts. To prevent a contract from becoming active due to a single large trade or an erroneous observation of the trade volume, a lower limit on open interest of 0.5% of the most actively traded contract is also placed.

The contract becomes inactive on the day following the last occurrence of reaching these limits. However, this decrease in activity indicates that most parties have already rolled their contracts, which leads to irregular price behaviour in the market. To roll simultaneously with market participants, we need to look at the First Notice Date (FND), the first date on which market participants could be notified that they have to physically deliver the commodity or receive its delivery. The dataset contains the last trading date (LTD) for all contracts. Based on the contract specifications as defined by the relevant exchange, the maximum difference between the LTD and the FND is established.[1] To mimic the rolling of contracts by market participants, we roll contracts one week before this date. Figure 2.1 shows for each commodity how this FND relates to the LTD, note here that the maximum difference is used, so that we are never late to roll any contracts. Furthermore, commodities with their shorthand written above the line denote those that are financially settled, the others are physically settled. The exact rolling procedure for each specific commodity is described in Appendix A.

## 2.2 Sector characteristics

While each commodity has its own specific supply and demand factors, broad patterns can be identified in the term structure of commodities belonging to the same sector. As these market

[1] Some contacts are settled financially rather than physically and thus have no first notice date. For these contracts, we define the first notice date as the first date on which information about the final settlement price is known.

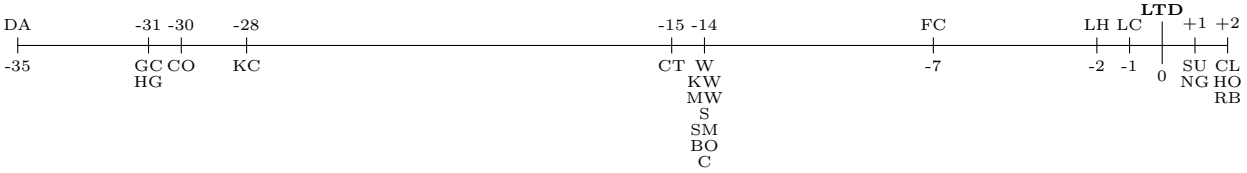
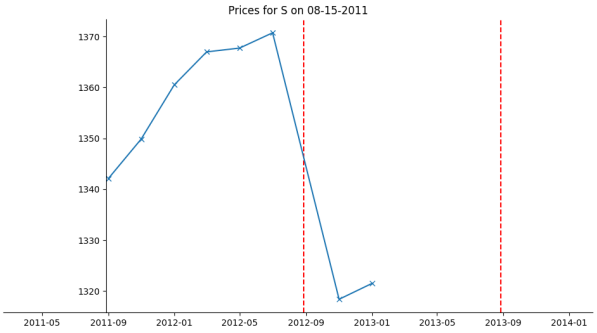



**Figure 1. Difference in days between the first notice date and last trading date.**

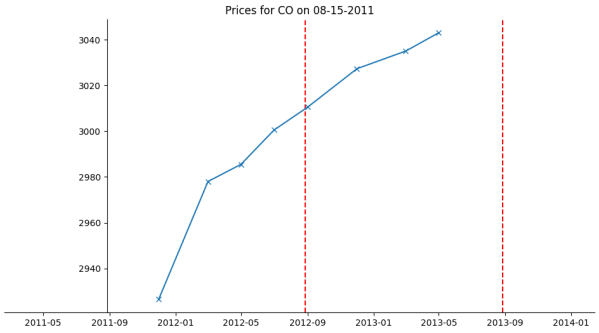
characteristics are important in analysing results, this section provides characteristics and shows examples of the term structure of markets belonging to the same sector. Firstly, agricultural commodities can be identified by their harvest-centred seasonal patterns, a term structure curve that is generally upward sloping, and around two years of actively traded contracts. Figure 2a shows an example of the soybean price curve that is typical for these agricultural markets. The y-axis and the red lines denote the end of August, in the middle of the soybean harvesting season. Indeed, the price curve is generally upward-sloping and the price declines after harvesting.

The main difference between agricultural commodities and soft commodities is the absence of the seasonal pattern for soft commodities. The softs are mostly grown around the equator, where the seasons are almost non-existent. Therefore, these commodities can be grown and harvested year round, leading to their futures price curves being mostly flat or upward sloping in normal situations. Figure 2b shows the term structure of cocoa on the same day and with the same vertical lines as Figure 2a for soybeans. The upward-sloping nature and the absence of seasonal effects are clearly visible here.

The energy commodities considered in this study are not as easy to generalise. Crude oil is a raw product from which both heating oil and gasoline are made, as well as a diverse set of consumer products. Natural gas is a separate raw product mainly used to warm homes in winter, which, as the name suggests, heating oil is also used for, but gasoline is not. The seasonal patterns in these energy commodities are mainly caused by seasonal demand, as they can be produced throughout the year. Therefore, the term structure curves for heating oil and natural gas are much alike, featuring strong seasonal patterns and a price peak in winter, when homes need to be warmed most. In addition, due to continuous demand, the price curve for crude oil futures is generally flat. The demand for gasoline peaks during the driving season, which is summer in the northern hemisphere, leading to a seasonal pattern in gasoline futures with a peak in summer. All energy commodities have in common that there are many futures contracts actively being traded, also further out on the curve. Figure 2c gives an example of the typical term structure curve for natural gas, where the vertical lines now represent the last day of December.

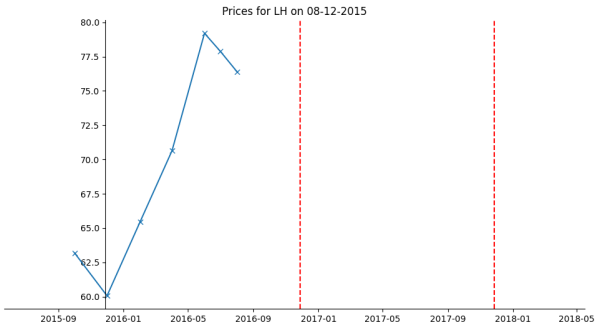



(a) **Agries**

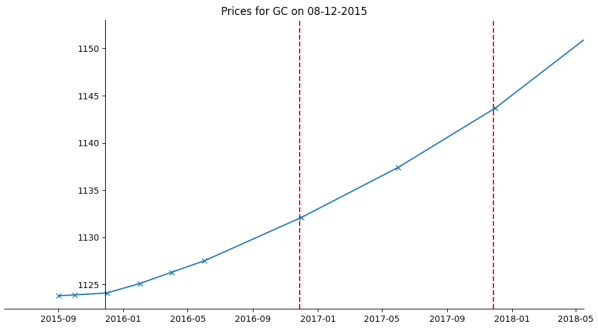



(b) **Softs**


(c) **Energy**




(d) **Meats**




(e) **Metals**

**Figure 2. Example term structures for each commodity sector**

Livestock futures are a complex sector characterised by influences from both the agricultural sector through feed prices and the energy sector through heating prices. This leads to complex observed term structures, both upward and downward sloping, often with a seasonal pattern of lowest prices just before winter, as livestock are slaughtered before winter to save on feed and heating costs. Figure 2d gives an example. Unlike livestock, metals are produced year round, easy to store, and have constant demand. Furthermore, any money spent holding metal cannot be used to earn the risk-free rate, making a futures contract more attractive than holding physical metal. Therefore, the price curve for metal futures is often slightly exponentially upward-sloping to represent the effect of compounded returns, as seen in Figure 2e. Compared to the copper curve, the gold curve is even flatter, as gold is mostly a financial commodity, so demand is not even dependent on world production.

# 3 Methodology

This section details the models and experimental setup used to answer the research questions. Section 3.1 describes the models and methods used to predict the term structure curves, while Section 3.2 explains how these predictions can be used to generate volatility predictions.

## 3.1 Predicting the term structure

This part of the methodology describes how the term structure curves can be modelled and predicted; it is divided into three distinct parts. First, Section 3.1.1 will detail how observed prices can be translated into a term structure curve for each commodity. Thereafter, Section 3.1.2 describes how we can model those curves in the time series dimension, and thus also how they can be used to generate predictions. Subsection 3.1.3 delves into the practical aspects of modelling term structure curves, describing the exact procedures used for estimating the models. Finally, Section 3.1.4 describes the experimental setup to create OOS predictions, as well as the criteria used to evaluate the predictions.

### 3.1.1 Price to curve

The very first step in predicting the term-structure curve of a commodity future is translating the observed prices into an unobserved price curve. To achieve this, our starting point is the Nelson-Siegel model, as given in Equation 1.

$$y_t(\tau) = \beta_1 + \beta_2 \left( \frac{1 - e^{-\lambda\tau}}{\lambda\tau} \right) + \beta_3 \left( \frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right) \tag{1}$$

where $y_t(\tau)$ denotes the yield at time $t$ of a bond with time to maturity $\tau$ in days. $\lambda$ is the decay and will be explained in more detail later. Finally, the betas are estimated coefficients. All bonds have the same loading on $\beta_1$, regardless of their maturity. Diebold and Li (2006) show that $\beta_1$ can thus be seen as a time-varying level factor. The loading on $\beta_2$ approaches one for bonds with a time to maturity that approaches zero and slowly declines to zero for longer-dated bonds. This coefficient can thus be seen as the premium for shorter dated bonds, corresponding to a time-varying slope factor.[2] Finally, the loading on $\beta_3$ is zero for very low time to maturity, reaches its maximum at some maturity $\tau^*$, and then declines to zero again for very long maturities. This allows the model to fit a variety of hump-shaped curves and can be interpreted as a time-varying curvature factor. Decay $\lambda$ controls how quickly slope and

[2]Depending on the sign of $\beta_2$ this 'premium' could be both positive and negative.

curvature loadings decline to zero and maturity $\tau^*$ at which the curvature loading reaches its maximum. Specifically, $\tau^*$ is approximately equal to $1.79328 * \lambda^{-1}$.[3]

Since bond yields are simply the inverse of bond prices and a bond can be seen as a prepaid futures contract on cash, the idea of modelling commodity futures prices using this same model makes intuitive sense. However, unlike bonds, commodity markets are influenced by a range of real-world dynamics, such as seasonality. For example, agricultural commodities can only be harvested during certain months, energy demand is highest during the winter, and livestock is often slaughtered before winter to save on heating and feeding costs. Failure to model these potential seasonal effects would result in a biased model. Since most of these effects are annual and following previous research by Karstanje et al. (2017), we choose to model the seasonal effects by adding a cosine wave with a periodicity of exactly one year. Therefore, the version of the Nelson-Siegel model that can be applied to commodity futures markets is given by Equation 2.

$$P_t(\tau) = L_t + S_t \left( \frac{1 - e^{-\lambda\tau}}{\lambda\tau} \right) + C_t \left( \frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right) + \kappa \cos\left(\omega\left(\tau + d_t\right) + \omega\theta\right) \tag{2}$$

where $P_t(\tau)$ denotes the price at time t of a future with time to maturity $\tau$ in days. $L_t$, $S_t$, and $C_t$ denote the time-varying factors which can be interpreted respectively as level, slope, and curvature. We assume that both strength of the seasonal effect $\kappa$ and the timing of the most expensive day in the year $\theta$ stay constant throughout the sample, so these cannot be interpreted as time-varying factors. $d_t$ denotes the which day of the year it is at time t, such that $\tau + d_t$ gives the number of days that have passed within the year at expiration of the contract. Finally, $\omega$ is a scalar equal to $\frac{2}{365}\pi$ ensuring that the periodicity of the cosine wave is one year.

### 3.1.2 Curve to curve

The previous section showed how to model the price curve of a given commodity on a given day. However, to achieve our goal of predicting the term structure, we need to think about how one curve relates to the next. To model the time series dimension of our problem, two approaches are used. The most simple is the day-by-day approach, where the model is estimated each day with the goal of achieving optimal fit on that given day. The next curve is then expected to be equal to this curve, based on the efficient markets premise that the latest price curve contains all relevant information for predicting future prices. Stated very simply in mathematical terms

[3]This follows from taking the derivative of the curvature loading with respect to $\tau$, setting equal to zero gives the trivial solution, $\lambda\tau^* = 0$ and the numerical solution presented, $\lambda\tau^* = 1.79328$

$P_{t+1}(\tau) = P_t(\tau)$.[4]

The second approach is based on the framework provided by Diebold et al. (2006), who show that the Nelson-Siegel model can be cast into a state-space representation if autoregressive properties are assumed on the level, slope, and curvature factors. The measurement equation of the state-space formulation is simply the multivariate version of 2, given by Equation 3.

$$\begin{pmatrix} P_t(\tau_1) \\ P_t(\tau_2) \\ \vdots \\ P_t(\tau_n) \end{pmatrix} = \begin{pmatrix} 1 & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} - e^{-\lambda\tau_1} \\ 1 & \frac{1-e^{-\lambda\tau_2}}{\lambda\tau_2} & \frac{1-e^{-\lambda\tau_2}}{\lambda\tau_2} - e^{-\lambda\tau_2} \\ \vdots & \vdots & \vdots \\ 1 & \frac{1-e^{-\lambda\tau_n}}{\lambda\tau_n} & \frac{1-e^{-\lambda\tau_3}}{\lambda\tau_3} - e^{-\lambda\tau_3} \end{pmatrix} \begin{pmatrix} L_t \\ S_t \\ C_t \end{pmatrix} + \kappa \begin{pmatrix} \cos\left(\omega\left(\tau_1 + d_t\right) + \omega\theta\right) \\ \cos\left(\omega\left(\tau_2 + d_t\right) + \omega\theta\right) \\ \vdots \\ \cos\left(\omega\left(\tau_n + d_t\right) + \omega\theta\right) \end{pmatrix} \tag{3}$$

Following the previous literature, we assume AR(1) dynamics on the state variables. Furthermore, like Karstanje et al. (2017) we assume the level factor to be nonstationary and model it as a first difference. However, this paper goes one step further by also assuming the slope and curvature factors to be non-stationary. The assumption is based on the results of Bianchi et al. (2023), which show that a momentum trading strategy on both slope and curvature factors has a highly significant Sharpe ratio, indicating the benefits of modelling autocorrelation in their changes rather than their absolute levels. Equation 4 shows the transition equation derived from these assumptions.

$$\begin{pmatrix} L_{t+1} - L_t \\ S_{t+1} - S_t \\ C_{t+1} - C_t \end{pmatrix} = \begin{pmatrix} \phi_1 & 0 & 0 \\ 0 & \phi_2 & 0 \\ 0 & 0 & \phi_3 \end{pmatrix} \begin{pmatrix} L_t - L_{t-1} \\ S_t - S_{t-1} \\ C_t - C_{t-1} \end{pmatrix} \tag{4}$$

where $\phi_1$, $\phi_2$, and $\phi_3$ denote the autocorrelation in changes in the level, slope, and curvature factors, respectively. Note that the matrix containing autoregressive coefficients is restricted to a diagonal form. This assumption promotes model parsimony, which has been shown, for example, by Christensen et al. (2011) to improve forecasting performance and estimation times.

### 3.1.3 Estimation

This Section describes in detail the exact specifications and procedures used to estimate the models, aiming to allow future researchers to exactly recreate the results presented in this study. First, we will discuss the estimation of the day-by-day model and thereafter the full state-space model. One of the advantages of the day-by-day model is the ease and speed with which it can

[4]Note that this is different from predicting the price of a certain contract to stay unchanged, since the same contract will have $\tau - 1$ time to maturity at $t + 1$

be estimated. The beta parameters in the standard form of the Nelson-Siegel model given in 1 can be estimated by Ordinary Least Squares (OLS) for fixed values of $\lambda$. The estimation of the seasonality adjusted version given in 2 is slightly more involved because of the non-linear effect of the $\theta$ parameter. The previous literature is divided on the optimal approach to determine the decay parameter $\lambda$. To accommodate the different market dynamics of all commodities, this article opts to view the decay $\lambda$ as an additional model parameter that must be optimised.

A two-pass procedure offers a simple solution for estimating both $\theta$ and $\lambda$. The model can be estimated for each value $\theta \in \mathbb{N}_{364}$, denoting the natural numbers from 0 up to and including 364, and twenty values of $\lambda$ on a logarithmic scale from 0.01 to 0.2, then the model with the highest R-squared is chosen. Additionally, the strength of the seasonal effect $\kappa$ is restricted to include only nonnegative values such that $\theta$ can always be interpreted as the most expensive day of the year.

As with most state-space models, we rely on the Kalman filter to estimate the state-space version of the model. To start, the model is cast into matrix notation, where Equation 5 denotes the transition equation, and Equation 6 denotes the measurement equation.

$$\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t = \mathbf{F}\left(\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t-1}\right) + \mathbf{Q} \tag{5}$$

$$\mathbf{P}_t = \mathbf{H}_t\,\boldsymbol{\xi}_t + \kappa\,\mathbf{X}_t + \mathbf{R}_t \tag{6}$$

In these equations $\mathbf{Q}$ and $\mathbf{R}$ correspond to the assumed error covariance structures. We assume $\mathbf{Q}$ and $\mathbf{R}$ to be independent, essentially imposing that the error terms in the measurement and transition equations are unrelated. Furthermore, considering the strong results in Christensen et al. (2011), we decide to limit both $\mathbf{Q}$ and $\mathbf{R}$ to diagonal matrices. This assumption means that there is neither covariance between the errors of the state variables, nor covariance between the errors in the estimated price curve. Both assumptions increase model parsimony, which has been shown to improve performance. Lastly, the diagonal elements of $\mathbf{Q}$ can be distinct, allowing the errors of the state variables to have heteroskedasticity. For $\mathbf{R}$ we assume that the errors for small time to maturity have the same distribution as the errors further away.

In order to facilitate estimation, the transition equation is first rewritten in a form where the unobserved states $\boldsymbol{\xi}_t$ can be estimated directly. This form is given by Equation 7.

$$\boldsymbol{\xi}_{t+1} = (\mathbf{I}_3 + \mathbf{F})\,\boldsymbol{\xi}_t - \mathbf{F}\,\boldsymbol{\xi}_{t-1} + \mathbf{Q} \tag{7}$$

Application of the Kalman filter then produces the predicted states and predicted covariance

matrix of the states as:

$$\boldsymbol{\xi}_{t|t-1} = (\mathbf{I}_3 + \mathbf{F})\,\boldsymbol{\xi}_{t-1|t-1} - \mathbf{F}\,\boldsymbol{\xi}_{t-2|t-2} \tag{8}$$

$$\boldsymbol{\Sigma}_{t|t-1} = \mathbf{F}\,\boldsymbol{\Sigma}_{t-1|t-1}\,\mathbf{F}^{\intercal} + \mathbf{Q} \tag{9}$$

In the Kalman updating step, the predicted values are updated according to the observed prices. This updating step is given by:

$$\boldsymbol{\xi}_{t|t} = \boldsymbol{\xi}_{t|t-1} + \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^{\intercal}\left(\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^{\intercal} + \mathbf{R}_t\right)^{-1}\left(\mathbf{P}_t - \mathbf{H}_t\boldsymbol{\xi}_{t|t-1} - \kappa\mathbf{X}_t\right) \tag{10}$$

$$\boldsymbol{\Sigma}_{t|t} = \boldsymbol{\Sigma}_{t|t-1} - \boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^{\intercal}\left(\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^{\intercal} + \mathbf{R}_t\right)^{-1}\mathbf{H}_t\boldsymbol{\Sigma}_{t|t-1} \tag{11}$$

We now have all the ingredients required to estimate the model by maximum likelihood estimation. The decomposition of the logarithmic likelihood prediction error is given by Equation 12.

$$\frac{1}{2}\sum_{t=1}^{T}\Big(-N_t\log(2\pi) - \log\left(\det\left(\mathbf{H}_t\,\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^{\intercal} + \mathbf{R}_t\right)\right) - \tag{12}$$

$$\left(\mathbf{P}_t - \mathbf{H}_t\boldsymbol{\xi}_{t|t-1} - \kappa\mathbf{X}_t\right)^{\intercal}\left(\mathbf{H}_t\,\boldsymbol{\Sigma}_{t|t-1}\mathbf{H}_t^{\intercal} + \mathbf{R}_t\right)\left(\mathbf{P}_t - \mathbf{H}_t\boldsymbol{\xi}_{t|t-1} - \kappa\mathbf{X}_t\right)\Big) \tag{13}$$

The extensive dataset in this paper, in combination with myriad time-varying components in the state-space model, makes the optimisation of this likelihood a nontrivial task. Therefore, a version of the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) which can be performed in parallel is implemented.

### 3.1.4 Prediction and evaluation

To create one-day-ahead term-structure predictions, the data needs to be split into a train and a test set. Because market conditions change over time, a moving rather than expanding window is chosen. Furthermore, the non-trivial estimation procedure of the state-space model makes it practically impossible to re-estimate the model after every prediction. Therefore, this paper opts for a procedure in which the model is estimated over a period of 10 years, after which one-day-ahead predictions are made for the 5 years thereafter, simply using the transition equation given by Equation 7. The model is then estimated on years 5 through 15, after which predictions are made for the five years thereafter. This procedure is repeated until predictions have been made for all days excluding the initial 10-year estimation window.

This approach leads to different parameter estimates in each estimation period for every commodity. To keep the reported estimates orderly and easily interpretable, a way to represent

all estimates using a single number is necessary. The arithmetic mean seems an obvious solution; however, a problem is encountered when applying this to the estimated $\theta$ parameters. Consider a situation where $\theta$, the most expensive day of the year, is estimated in one period to be 360, 364 in the next, and 14 in the last period. To any human reader, it is obvious that the 'average' most expensive day should be somewhere at the end of December or beginning of January. However, the arithmetic mean is 246, corresponding to a day in August. The median provides some relief but is still not the desired solution. To solve the problem, we use the circular mean, a concept designed to solve a similar problem of calculating the average angle from a set of angles. The definition of the circular mean can be found in Mardia et al. (2000), which we use directly after transforming $\theta$ from days to degrees simply by dividing 360 by $\theta$.

To evaluate the predictions made by the daily and state-space versions of the models, this paper uses the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean squared error (RMSE); they are defined according to Equations 14, 15, and 16, respectively. The MAPE has the advantage of being insensitive to the price of the commodity, allowing for cross-commodity comparisons. Moreover, it is not influenced by changes in price level throughout the 50-year sample, while both the MAE and the RMSE will overweight errors made in periods with higher prices. This is because a 1% error when the price is 100 is half as large as a 1% error when the price is 200. Because the errors are squared before the average is taken in the RMSE, the effect is even stronger when using this metric. This paper aims to investigate to what degree the term-structure curve can be predicted, regardless of the price level. Therefore, MAPE is the preferred evaluation metric.

$$\text{MAE} = \frac{1}{T-s}\sum_{t=s}^{T}\sum_{i=1}^{n_t}\frac{1}{n_t}|P_t(\tau_i) - \widehat{P_t(\tau_i)}| \tag{14}$$

$$\text{MAPE} = \frac{1}{T-s}\sum_{t=s}^{T}\sum_{i=1}^{n_t}\frac{|P_t(\tau_i) - \widehat{P_t(\tau_i)}|}{P_t(\tau_i)} \tag{15}$$

$$\text{RMSE} = \sqrt{\frac{1}{T-s}\sum_{t=s}^{T}\sum_{i=1}^{n_t}\frac{1}{n_t}\left(P_t(\tau_i) - \widehat{P_t(\tau_i)}\right)^2} \tag{16}$$

Where $s$ denotes the first day after the first estimation window of the state-space model, which corresponds to 10 years after the start of the full dataset. Furthermore, $n_t$ indicates the number of active contracts on day $t$.

Classically, testing whether the predictions of one model are statistically better than those of another is done using the Diebold-Mariano test (Diebold & Mariano, 2002). The null hypothesis in this test is that the average loss of the predictions of both models is equal, where the loss

of a prediction is often defined as the squared or absolute error. Therefore, rejecting the null hypothesis using a two-sided confidence interval indicates that the predictions from one model are significantly better than those from the other. However, the classic specification of the Diebold-Mariano test is not applicable here, as there are multiple futures traded on each day. To solve this problem we can turn to Qu et al. (2023), who extend the Diebold-Mariano test statistic to apply to balanced panel data. Even so, in our data the number of active futures can vary from day to day, so we slightly adapt their statistic to fit unbalanced panel data as well. The final form is given by Equations 17, 18, and 19.

$$\Delta L_{i,t|t-1} = \left(P_t(\tau_i) - \widehat{P_{t,m_1}(\tau_i)}\right)^2 - \left(P_t(\tau_i) - \widehat{P_{t,m_2}(\tau_i)}\right)^2 \tag{17}$$

$$J_{n,T}^{DM} = T^{-\frac{1}{2}} \frac{\sum_{t=s}^{T} \sum_{i=1}^{n_t} n_t^{-\frac{1}{2}} \Delta L_{i,t|t-1}}{\hat{\sigma}\left(\Delta L_{t|t-1}\right)} \tag{18}$$

$$\hat{\sigma}\left(\Delta L_{t|t-1}\right) = \sqrt{\sum_{j=-J}^{j=J} \left(1 - \frac{j}{J}\right) \hat{\gamma_h}(j)} \tag{19}$$

Where the subscripts $m_1$ and $m_2$ denote by which model the predictions, indicated by a hat, have been made. Furthermore, $\hat{\gamma_h}(j)$ denotes the j-th order autocorrelation of the forecasts with the forecast horizon $h$, the exact definition of which can be found in Qu et al. (2023). In this paper, the forecast horizon $h$ is equal to 1 day and the maximum lag length $J$ is set to 20.

### 3.2 Translating to volatility predictions

This Section introduces the methodology used to evaluate to what degree the term-structure predictions generated in 3.1.4 can be used to improve volatility predictions for the underlying commodity market. A baseline volatility model is augmented using various measures derived from the predicted factors. Furthermore, we investigate whether some standard machine learning models are capable of increasing forecast performance.

The baseline model used is the standard heterogeneous autoregressive (HAR) model as introduced in Corsi (2009). However, due to limited data availability in the full sample period, it is not possible to use the realised volatility (RV) as a measure of volatility. Instead, this paper defines the volatility on a given day as the true range (TR) of the most nearby futures contract. The true range on day $t$ is defined as the maximum difference between the closing price on day $t-1$ and the highest and lowest price on day $t$. Therefore, it also accounts for any possible price movements between the closing price on day $t-1$ and the opening price on day $t$. This measure of volatility has been used only sparsely in academic literature, but is popular

with practitioners in financial markets. Our baseline version of the HAR model is thus defined as

$$TR_t = \beta_0 + \beta_1 TR^d_{t-1} + \beta_2 TR^w_{t-1} + \beta_3 TR^m_{t-1} + u_t \tag{20}$$

Where $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ are unknown parameters that can be estimated using OLS. $TR^d_{t-1}$, $TR^w_{t-1}$, $TR^m_{t-1}$ denote the lagged daily, weekly, and monthly true range volatility estimates defined as $TR^d_{t-1} = TR_{t-1}$, $TR^w_{t-1} = \frac{1}{5}\sum_{i=1}^{5} TR_{t-i}$, and $TR^m_{t-1} = \frac{1}{22}\sum_{i=1}^{22} TR_{t-i}$.

To investigate whether the forecasting performance of this baseline model can be improved through measures derived from the predicted price curves, we consider 14 possible explanatory variables. Each is added to Equation 20 separately and estimated by OLS. Furthermore, the whole set of 14 is added to a single model which is then estimated by OLS and LASSO, a regularised version of OLS. Lastly, a basic neural network and random forest are applied to the same set of regressors.

### 3.2.1 Description of exogeneous regressors

The 14 possible explanatory variables can be divided into four groups, all of which are described here. Firstly, we have those based on the absolute value of the slope and curvature coefficients. These are designed with the idea that a very strong contango or backwardation structure indicates a distressed market, which would be a sign of higher volatility. To reduce the sensitivity to outliers, these are defined as dummy variables equal to one when the slope or curvature coefficient is further away from zero than a certain threshold and zero otherwise.[5] Below gives the definition for the Absolute Slope Dummy (ASD) based on the estimates by the daily-estimated model, the others are defined in the same way.

$$ASD^{daily}_t = \begin{cases} 1, & \text{if } \zeta^{low}_{S,daily} < \frac{|S^{daily}_t|}{\bar{P}_t(\tau_i)} < \zeta^{high}_{S,daily} \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

where $S^{daily}_t$ denotes the slope factor on day $t$ estimated by the daily model. We divide by the mean price to be able to compare across commodities. The upper bound for each of the variables is given by $\zeta^{high}$, and the lower bound is given by $\zeta^{low}$. Both thresholds are parameters to be optimised.

Since the 'normal' shape of the price curve can differ from commodity to commodity, it is important that alongside absolute values of the estimated slope and curvature factors, there are also regressors implemented based on the relative difference between the current slope and

[5] As will be shown in Section 4, sometimes estimated slope and/or curvature factors are unreasonably large, which forces the use of a maximum threshold on the deviation as well as a minimum.

curvature and the 'normal' slope and curvature. The in-sample median slope or curvature is chosen to represent the 'normal' state, since it is less sensitive to large outliers. Similarly to Equation 21, a dummy representation is implemented. Moreover, the direction and magnitude of the deviation might hold important information, so a version is implemented without the dummy structure. Equation 22 provides the specification of the Relative Slope Dummy (RSD) for the daily model and Equation 23 shows how the Relative Slope Coefficient (RSC) is defined for the daily model. The state-space and curvature versions are defined similarly.

$$RSD_t^{daily} = \begin{cases} 1, & \text{if } \hat{\zeta}_{S,daily}^{low} < \frac{|S_t^{daily} - median_t\left(S^{daily}\right)|}{\overline{P_t}(\tau_i)} < \hat{\zeta}_{S,daily}^{high} \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

$$RSC_t^{daily} = \frac{S_t^{daily} - median_t\left(S^{daily}\right)}{\overline{P_t}(\tau_i)} \tag{23}$$

The last class of explanatory variables is based on the prediction errors made by the state-space and daily estimated models. The hypothesis here is that our curve predictions are less accurate when the curve evolves differently from normal, which would indicate a changed, and possibly more volatile, market environment. We consider the possibility that only the magnitude of the errors contains predictive values by implementing the mean absolute error on a given day $t$ as a predictor of the volatility on day $t+1$. Additionally, by simply implementing the mean error on day $t$ as a predictor, the possibility is also considered where the direction of the error contains predictive information. The variant based on the mean absolute error (MAE) is given in Equation 24 and the one based on the mean model error (MME) is given in Equation 25.

$$MAE_t^{daily} = \frac{1}{n_t}\sum_{i=1}^{n_t} |P_t(\tau_i) - \widehat{P_t(\tau_i)_{daily}}| \tag{24}$$

$$MME_t^{daily} = \frac{1}{n_t}\sum_{i=1}^{n_t} P_t(\tau_i) - \widehat{P_t(\tau_i)_{daily}} \tag{25}$$

Three machine learning methods are employed to investigate whether their ability to allow non-linear effects increases forecasting performance. The LASSO model by Tibshirani (1996) is a regularised regression that allows the coefficients of some potentially uninformative predictors to be set to zero. Furthermore, a basic version of the neural network with one hidden layer is implemented using standard hyperparameter configurations. Lastly, a standard implementation of the random forest is employed (Breiman, 2001).

### 3.2.2 Estimation and prediction

All regressors described in Subsection 3.2.1 are implemented both for the state-space estimation method and the daily estimated method. Furthermore, where applicable, they are applied to both slope and curvature factors. This will give us a total of 20 additional variables that could have predictive power. To test which is strongest, all are incorporated separately as an additional explanatory variable in Equation 20. To investigate whether there exist interaction effects or whether some variables draw out the effects of others, some specifications containing all the new explanatory variables and the baseline HAR ones are considered. The simplest form is given by the baseline Equation 20 with an additional 20 explanatory variables. This model is estimated through OLS, as well as using the LASSO regression, one-layer neural network (NN), and the random forest (RF). As this is meant as an initial investigation rather than an exercise to maximise forecast accuracy, the hyperparameters are not optimised. Their initial values can be found in Appendix C.

All volatility forecasting is done out-of-sample. Since financial markets are characterised by changing relations, a rolling window approach is used. Firstly, all models are estimated over a period of 500 observations; since there are about 250 business days in a year, this is approximately equal to a period of two years. Thereafter, one-day-ahead and one-month-ahead volatility predictions are made for the following 125 days.

All models are evaluated by the MAE of their volatility predictions. For easy comparison, these are reported as a fraction of the MAE of the baseline model for that commodity. Furthermore, we calculate the standard Diebold-Mariano test as in Diebold and Mariano (2002) for each model compared to the baseline model. The null hypothesis of this test is that the average absolute error of both models is equal, and rejection indicates that the predictions made by one model have significantly lower MAE than those made by the other model.

# 4 Results

In this Section results are presented to show to what degree it is possible to predict volatility in commodity markets using the shape of predicted term structure curves. Initially, the accuracy of one-day-ahead price curve predictions is evaluated. The daily method of estimating the Nelson-Siegel type model is compared to the state-space version both quantitatively through a large out-of-sample (OOS) test and graphically through an inspection of the predicted curves. Thereafter, it is investigated whether these curve predictions are capable of improving predictions from a standard HAR volatility model by including shape estimates as additional parameters in the model.

## 4.1 Term structure predictions

This part focuses on the results of predicting the term structure curve using the daily-estimated and state-space Nelson-Siegel models. Initially, parameter estimates are discussed and compared across models, commodities, and time. Thereafter, the OOS predictive accuracy of the models is compared and the models are formally tested using a Diebold-Mariano test. Finally, a graphical investigation is conducted into the markets for which the predictions are worst and best.

### 4.1.1 Parameter estimates

Table 2 shows the estimated $\lambda$, $\kappa$, and $\theta$ for both estimation methods and all commodities, averaged over the entire sample period. The decay parameter $\lambda$ determines how quickly the loadings of the slope and curvature factors decline to zero. Considering the estimated values given in the first column, large differences between the values reported for the daily estimated model and the state-space version are evident. For the daily model, all values are of the same order of magnitude, with the exception of gold, ranging from 0.0135 for crude oil to 0.0702 for lean hogs. This is approximately equal to $\tau^*$, the maturity at which the curvature loading is maximised, ranging from 25 to 133 days for all these commodities. Figure 3a shows how the loadings of the slope and curvature factors change as a function of the time to maturity for the decays of crude oil and lean hogs. Slope and curvature factor loadings drop to zero much more rapidly for lean hogs than for crude oil. This implies that the prices of contracts further down the curve are mostly determined by the level factor and seasonality effect for lean hogs, while the slope and curvature factor influence a larger part of the curve for crude oil, ultimately leading to the price curve for lean hogs being only a cosine wave when time to maturity is high. The range of average estimated decays is much larger for the model estimated using the state-space methodology. Again, with the exclusion of gold, $\tau^*$ is between 9 (natural gas) and 285 days

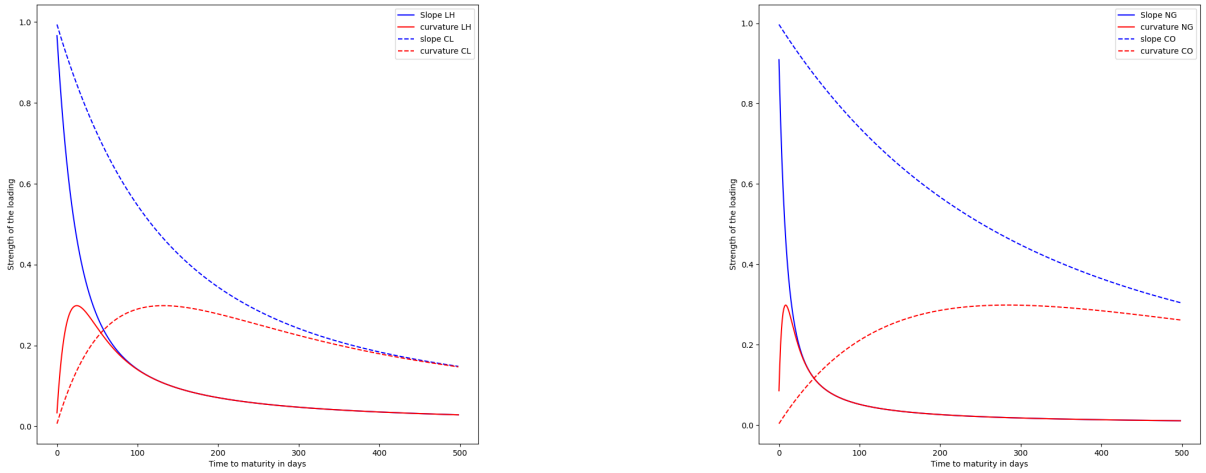
**Table 2. Average estimated parameters for both models**

| | $\lambda$ | | $\kappa$ | | $\theta$ | |
|---|---|---|---|---|---|---|
| | daily | state-space | daily | state-space | daily | state-space |
| W | 0.0279 | 0.0095 | 3.2189 | 2.0814 | 320.85 | 318.15 |
| KW | 0.0218 | 0.0128 | 2.8967 | 1.9329 | 333.30 | 311.84 |
| C | 0.0324 | 0.0476 | 2.4988 | 0.7308 | 200.93 | 275.77 |
| S | 0.0398 | 0.0084 | 2.9483 | 0.7657 | 197.83 | 274.69 |
| SM | 0.0418 | 0.0250 | 2.5789 | 0.4455 | 194.06 | 287.89 |
| BO | 0.0373 | 0.1448 | 1.2085 | 0.1222 | 202.41 | 297.71 |
| FC | 0.0621 | 0.1561 | 3.5866 | 0.2675 | 76.83 | 311.36 |
| LC | 0.0615 | 0.1353 | 3.5175 | 1.8589 | 296.75 | 296.88 |
| LH | 0.0702 | 0.1267 | 11.8872 | 3.0403 | 191.88 | 268.37 |
| CL | 0.0135 | 0.0583 | 0.3156 | 0.1663 | 31.79 | 314.39 |
| HO | 0.0251 | 0.0099 | 2.2307 | 1.6076 | 8.11 | 355.18 |
| RB | 0.0577 | 0.0204 | 5.8868 | 4.1208 | 208.35 | 240.91 |
| NG | 0.0216 | 0.1934 | 6.9786 | 17.8247 | 6.76 | 299.71 |
| GC | 0.0026 | 0.0014 | 0.1434 | 0.0607 | 308.73 | 309.51 |
| HG | 0.0213 | 0.1123 | 0.3744 | 0.0234 | 166.95 | 303.61 |
| KC | 0.0281 | 0.0813 | 0.7514 | 0.0791 | 163.07 | 315.88 |
| CT | 0.0369 | 0.1553 | 2.1363 | 0.7801 | 191.29 | 265.66 |
| SU | 0.0267 | 0.1778 | 1.3375 | 8.2699 | 318.53 | 303.34 |
| CO | 0.0251 | 0.0063 | 1.7105 | 0.0048 | 99.68 | 298.73 |

*Note.* Table gives the average estimated parameters for both models throughout the sample, for the daily model this corresponds to the average of all daily estimates, while for the state-space model it corresponds to the average of the estimation periods. $\kappa$ is reported as a percentage of the mean price. Furthermore, the cyclical mean, rather than the arithmetic mean is presented for $\theta$ because the average of some day in December and some day in February should be in January, not in July.

(cocoa). In Figure 3b, we see that this allows us to have different loadings on the slope and curvature factors for the entire cocoa curve. However, for natural gas, the loading on the slope and curvature factors are very similar from 30 days to maturity, and are both less than 10% from 50 days to maturity. This could bring adverse effects in predicting the curves, since any variations in the long end of the term-structure can only be ascribed to seasonality effects, or to a very small fraction of the slope and curvature factors. Therefore, such large estimated values $\lambda$ could indicate outsized slope and curvature factors or seasonality effects, which is investigated in Section 4.1.2.

Gold is the commodity with the lowest estimated decay for both models. This is not surprising, because gold is quite different from all other considered commodities. Whereas the main use of all other commodities is in production and consumption, gold is mainly used as an alternative to holding currency. Therefore, the main factor influencing the term structure of gold is the risk-free rate which would be earned while holding currency but not when holding gold. As seen in Figure 2e, this makes for an exponentially upward sloping term structure due to compounded rates. The higher the estimated decay is, the faster the effects of the slope and




**(a) Daily model** **(b) State-space model**

**Figure 3. Shape of the slope and curvature loadings for given decay values**

curvature factors will decline to zero, and thus the model will be unable to capture the upward sloping curve at long maturities. The decay needs to be at least low enough to capture the increasing trend from the second to last to the last actively traded futures contract, which is why the estimates by both models are so low.

The values reported for $\kappa$ in Table 2 can be interpreted as the percentage of the price of the commodity that is determined by seasonal effects. For crude oil, gold, copper, and coffee, both models estimate the strength of the seasonal effect to be less than 1%. This effect is so small that these commodities can be considered nonseasonal, matching fundamental insights about these markets. Therefore, the reported $\theta$ is not significant and should be ignored. In all other markets except for natural gas and sugar, the estimated strength of the seasonal effect is stronger for the daily model than for the state-space version. The two exceptions happen to also be the markets with the highest average estimated decay in the state-space model, providing support for the hypothesis in the previous paragraph that the high estimated $\lambda$ leads to an inflated estimate of the strength of the seasonality effect.

Our sample contains some highly similar commodities for which it might be interesting to compare seasonality estimates. First, wheat and kansas wheat are the futures contracts for soft red winter wheat and hard red winter wheat, respectively. Both are planted and harvested around the same time. Therefore, estimates of $\theta$, which denotes the most expensive day of the year, are expected to be similar. For both models, this expectation is satisfied, with the average estimate of the most expensive day of the year being around the 320th day. Intuitively, this makes sense as that is around the time winter wheat is planted, so it will take some time until a new harvest is available. The second interesting combination of commodities consists of

soybeans, soybean meal, and soybean oil, since soybeans are the raw product and both soybean meal and soybean oil are processed products. For these products, the state-space model is unable to detect significant seasonal effects, while the daily model estimates a seasonal effect for all three, but the strength in soybean oil is about half that of the other two. Based on the fact that soybeans are mostly harvested during one period of the year, a seasonal effect is expected in this market. Furthermore, the estimated $\theta$ of about 200 in the daily model is consistent with the harvesting periods of the largest soybean growers, the United States, and Brazil. The reduced strength of the seasonality effect in soybean oil might be due to it being less costly to store compared to soybeans and soybean meal.

### 4.1.2 Predictive accuracy

The previous section discussed the parameter estimates by both models, where it was concluded that the high estimated values for the decay $\lambda$ might lead to issues in estimating and predicting the price curves. This section examines the forecasting performance of the models to see if this is indeed the case. To evaluate the broad predictive ability of the models, three measures of average errors are used. Most importantly, the mean absolute percentage error (MAPE). Since it is standardised by the average price, this measure allows us to compare between models as well as commodities. Furthermore, it shares with the mean absolute error (MAE) an insensitivity to outliers, which is particularly useful in this research, since the prices of commodities tend to experience large price fluctuations during our 50-year sample, and we do not wish to overweight periods where the price was higher. Lastly, the RMSE is reported to evaluate how well the models deal with outliers. Furthermore, the difference between the RMSE and MAE provides a measure of the stability of the errors, since the larger this difference is, the more dispersed the errors are.

Table 3 reports lower MAPE, MAE, and RMSE for the daily-estimated version of the model for every commodity.[6] This indicates that the daily estimated model is better able to predict the price curve on a one-day prediction horizon than the state-space version. Using the pooled version of the Diebold-Mariano test, the null hypothesis of both models having equal predictive accuracy can be rejected at the 1% significance level for all commodities. The results are not different if the test is based on the MAE or MAPE, but note that it has not been proven that the pooled DM-test also holds true for loss functions other than the RMSE. However, depending on the commodity, the difference between the state-space and daily versions of the model is

[6]Note that the daily model requires only one day before predictions can be generated, while for the state-space model there is an estimation window of ten years. To achieve a fair comparison, only those days are taken into account for which both models have made a prediction

**Table 3. Price prediction errors over the entire sample**

| | MAPE | | MAE | | RMSE | |
|---|---|---|---|---|---|---|
| | daily | state-space | daily | state-space | daily | state-space |
| W | **1.179** | 1.946 | **6.139** | 9.814 | **10.638*** | 15.744 |
| KW | **1.194** | 1.994 | **6.675** | 10.443 | **11.740*** | 15.896 |
| C | **1.052** | 2.062 | **3.997** | 7.699 | **7.138*** | 12.139 |
| S | **1.006** | 1.500 | **8.608** | 12.678 | **14.523*** | 18.887 |
| SM | **1.082** | 1.864 | **2.751** | 4.597 | **4.382*** | 6.938 |
| BO | **1.043** | 3.957 | **0.331** | 1.029 | **0.515*** | 2.449 |
| FC | **0.653** | 1.232 | **0.781** | 1.490 | **1.318*** | 2.221 |
| LC | **0.800** | 1.731 | **0.779** | 1.663 | **1.430*** | 2.265 |
| LH | **1.783** | 5.274 | **1.166** | 3.497 | **2.351*** | 4.528 |
| CL | **1.250** | 3.833 | **0.722** | 2.055 | **1.114*** | 4.026 |
| HO | **1.229** | 4.313 | **1.874** | 4.966 | **3.004*** | 8.381 |
| RB | **1.647** | 2.944 | **2.671** | 4.627 | **3.896*** | 6.656 |
| NG | **2.640** | 14.239 | **0.149** | 0.697 | **0.222*** | 1.160 |
| GC | **0.754** | 1.031 | **13.692** | 18.744 | **19.195*** | 25.219 |
| HG | **1.048** | 1.819 | **1.861** | 3.067 | **3.806*** | 5.128 |
| KC | **1.391** | 2.358 | **1.867** | 3.137 | **2.825*** | 4.910 |
| CT | **0.987** | 2.348 | **0.701** | 1.656 | **1.484*** | 2.721 |
| SU | **1.174** | 12.011 | **0.173** | 1.134 | **0.261*** | 1.856 |
| CO | **1.193** | 1.374 | **22.932** | 26.328 | **32.121*** | 36.158 |

*Note.* Table reports the mean absolute percentage error (MAPE) as a percentage, mean absolute error (MAE) and root mean squared error (RMSE) for both models over all the predictions. The * indicates we reject the null hypothesis of equal average loss at the 1% level using the pooled Diebold-Mariano test statistic.

more or less pronounced. For example, for cocoa the MAE of the state-space model is only 1.1 times larger than that of the daily estimated model. On the contrary, for sugar, it is 10 times as large. Later, a graphical inspection of the curve predictions for both markets is performed to investigate what might be causing the large dispersion between commodities.

The daily-estimated model has been established to produce more accurate price curve predictions on the one-day prediction horizon. However, MAPE also gives us a way to compare predictive accuracy between commodities, allowing us to investigate whether some types of commodities are more predictable than others or whether the state-space model has a comparative advantage somewhere. First, both models have the worst performance in the natural gas market. This is not entirely unexpected as the natural gas market is notorious for its high volatility. Furthermore, the seasonality specification using a single cosine wave as given by Equation 2, is not completely accurate for the natural gas market. While most of the seasonal price fluctuation occurs due to increased demand for heating during the winter, there is a second, smaller, price increase in the middle of summer due to increased demand for cooling. Figure 4a shows the in-sample fit of the daily model on the natural gas market. It shows clearly that the single cosine wave as a measure of seasonality is unable to fit the small price increases in summer,

represented in the plot by the red crosses within the troughs of the seasonal pattern. To compensate, the model seems to underestimate the strength of the seasonal effect in winter, leading to undershooting the price peaks at the top of the seasonal pattern. The state-space version of the model has the additional issue of having a fixed strength of the seasonal effect, while prices are very volatile for natural gas and can change a lot in the 10-year estimation window. This means that the same seasonal effect could be 20% of the price in the first year and 50% in the tenth year. Further research might want to implement a seasonal effect that adapts to changes in the general price level or is time-varying in some other way.

On the other end of the commodity spectrum lies gold. For the state-space model, this is the best performing market and it is the second best for the daily model. This is also not very surprising, as gold can be seen as a currency and has very low volatility. Furthermore, gold is easily storeable, has no seasonalities, and shocks to supply or demand are very rare. All of these aspects lead to a stable term-structure curve which both models are able to handle well.

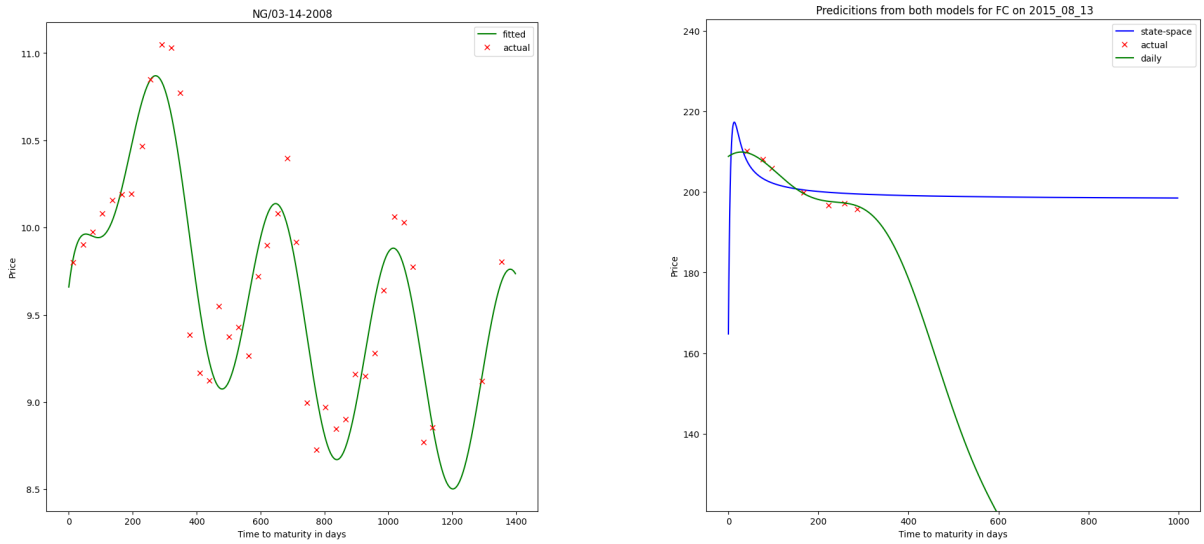
More interesting is the best performing market of the daily model, feeder cattle. Feeder cattle are cows that are not yet fully grown. Their price is influenced not only by the demand and supply of meat, but also the price of feed, housing, and heating. These factors introduce complex seasonality and volatility, much like the natural gas market. The difference in performance for these commodities can be explained by the difference in the number of contracts actively traded at any time. Figure 2 in Appendix B shows that there are usually more than 20 different contracts that are traded on natural gas, while for feeder cattle there are only 5 or 6. The daily model has 4 estimated coefficients, allowing it to fit any 4 price points perfectly in-sample.[7] Figure 4b shows a situation where the predicted curve by the daily model is almost exactly equal to the prices, allowing the daily model to achieve a very low MAE. The drawback is that the coefficients that perfectly fit the prices at a given point in time may not represent the actual hidden level, slope, and curvature factors. For example, in the figure shown, the estimated level factor is -2500, indicating the price of a contract with very long time to maturity would be -2500. In Section 4.2 this might cause problems, as the estimated slope and curvature factors are used in a volatility prediction model.

### 4.1.3 Graphical inspection of best and worst markets

In this section, we investigate why the relative performance of the state-space model compared to the daily model is so different between commodities.

First, sugar, the commodity where the state-space version has the worst relative perfor-

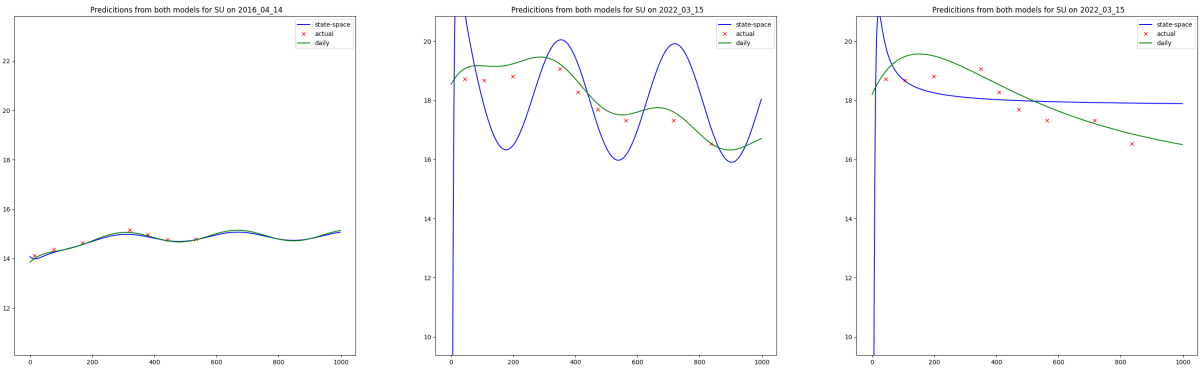
[7] Technically, the $\lambda$ and $\theta$ also have to be estimated, so there are 6 estimated coefficients. However, their non-linear nature does not necessarily allow for a perfect fit on 6 price points.




**(a) In-sample natural gas fit** **(b) Feeder cattle price curve predictions**

**Figure 4. Estimation issues in select markets**

mance, is investigated. Figure 5a shows a situation where the state-space model is not much worse than the daily estimated model, and Figure 5b shows one in the next estimation window where it is. When comparing the blue lines in both pictures, two things stand out in Figure 5b. The very pronounced seasonal cycles and the almost vertical price curve for a very short time to expiration. Although it may not be obvious immediately, these issues are likely related. First, note that the sugar contracts are rolled 9 days before expiration, so that there will never be any observations on the very short end of the curve. Therefore, the vertical line has only a limited impact on the predictive performance of the model. Figure 5c shows the same predictions as Figure 5b, but with the seasonal effect removed. We notice how using only the level, slope, and curvature factors the daily estimated model is still able to predict the general shape of the curve well, the state-space version, on the other hand, predicts just a horizontal line for most of the curve. Now it seems that the seasonal effect is estimated to be so strong that it compensates for the horizontal line not being able to follow the natural backwardation of the sugar market. The vertical and horizontal lines could be explained by a strong curvature factor combined with a high decay, as this would translate to a very low $\tau^*$. Indeed, the estimated decay coefficient in this period is equal to 0.19, leading to the loading on the curvature factor reaching its maximum at only 8 days to maturity, which explains the peak we see around this period. A possible explanation for the tendency of the model to predict a flat line rather than following the shape of the market could be that frequent jumps in the shape of the term structure during the estimation period occur, such that simply predicting the same level for all maturities leads to lower overall

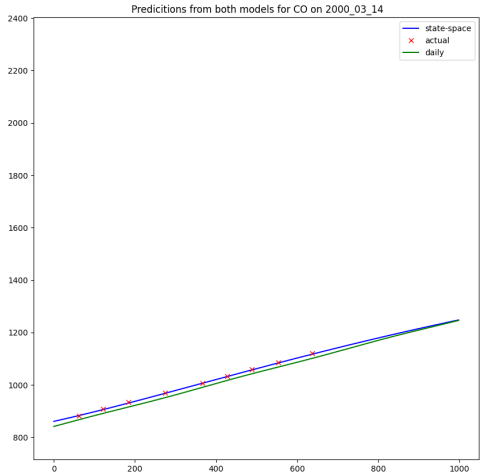



**(a) Good relative prediction** **(b) Bad relative prediction** **(c) Seasonal effect removed**

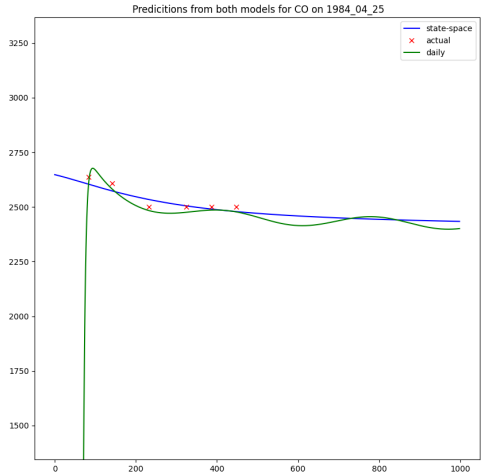
**Figure 5. Accurate and non-accurate predictions for sugar**

error than trying to predict the actual shape. If this is the case, we would also expect the daily estimated model to sometimes make really big mistakes while making small mistakes most of the time. This hypothesis can easily be checked by looking at the spread between the MAE and the RMSE for the daily estimated model. As mentioned before, the larger the spread, the more dispersion in the size of the errors, so the more likely there occur big jumps in the prediction period. Interestingly, this is not the case, since the MAE is about 0.75 times the RMSE in this specific period, while it is 0.65 times the RMSE in the full sample.

Perhaps a deeper look into the state-space model's relative best performing market, cocoa, will provide more insight. Cocoa grows only in a narrow strip around the equator worldwide, in this area the seasons are much less pronounced or even non-existent. Therefore, the term structure of cocoa generally has a very stable structure that looks like a straight line, possibly slightly upward or downward sloping, and in some cases with a slight curve. Figure 6 shows some examples. For both models, this provides a low-noise environment where their predictions are quite good. It is in this specific environment where the power of modelling the changes in state variables through the state-space methodology becomes visible. Since for the daily model $\widehat{P_{t+1}}(\tau) = P_t(\tau)$, we can safely assume that the day before Figure 6a the prices were on the green line in the figure. The daily estimated model simply predicts the prices to be on that line again and is therefore slightly below the actual price curve. However, the state-space model predicts the change based on the autoregressive properties of the level, slope, and curvature factors, allowing it to predict the prices in 6a almost perfectly. So why does the state-space model still perform worse than the daily model in this market? Figure 6b shows the prices of the first two active contracts being slightly elevated relative to all others. For the cocoa market, this is an extraordinary situation, possibly caused by some short-term supply or demand factors. Taking




**(a) Standard situation**



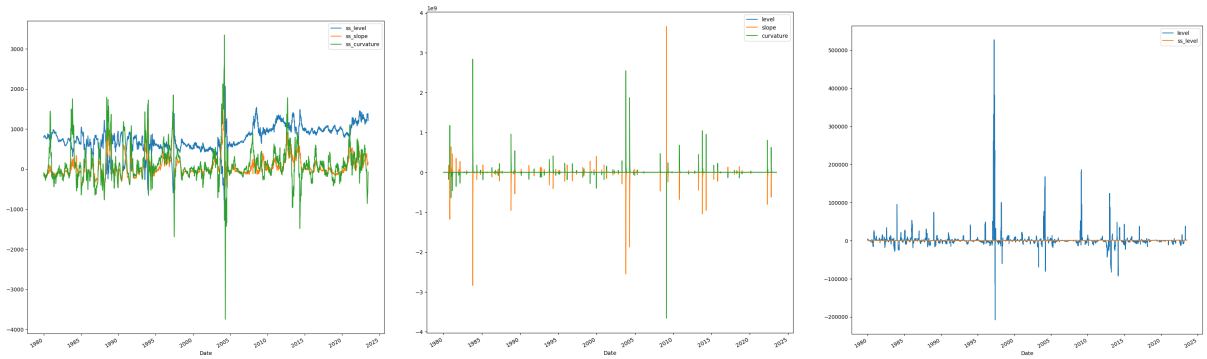

**(b) Special situation**

**Figure 6. Some cocoa market term-structures**

into account only the actual prices given by the crosses, the prediction of the curve by the daily model is better, leading to a lower MAE and RMSE on this day. However, if we consider the entire curve, the state-space model seems to provide a much more realistic and robust estimate of what the shape could be like. Therefore, it is likely that any inference about the market made by inspecting the predicted curves will be better using the state-space version. This hypothesis is evaluated in Section 4.2.

The instability of the factor estimates by the daily model is thus a potential threat to the performance of the volatility prediction models. To investigate the severity of the problem, predictions made by the daily estimated and state-space versions of the model can be plotted through time and compared. Figure 7a gives the level, slope, and curvature factors estimated using the state-space methodology for the soybean market. Figure 7b shows the same thing, but with the factors estimated using the daily estimation methodology. The two look nothing alike, since Figure 7b is completely dominated by a few very extreme estimates of the slope and curvature factor. The severity of the problem is fully apparent when considering the scaling of the y-axis in both pictures. Whereas the maximum estimated slope in the whole sample for the state-space model is about 3 thousand, it is more than 3 billion using the daily methodology. As we hypothesised while discussing Figure 4b, the extreme estimate for the slope factor is always compensated by an equally extreme but opposite in sign estimate of the curvature factor, combined with high decay, this can still provide accurate price predictions. However, it will likely lead to problems when factors are used to predict volatility.

In contrast, the factor estimates produced by the state-space model seem very reasonable, with the estimated level factor always around the general price level of soybean futures contracts.



**(a) State-space predictions** **(b) Daily factor predictions** **(c) Level for both models**

**Figure 7. Comparison of predicted factors for the soybean market**

To show the contrast in stability of the estimates, Figure 7c plots the state-space level factor and the daily level factor on one graph. Note that the level factor is not even visible in Figure 7b, as it is dwarfed by the estimated slope and curvature. Yet still, the level factor estimated by the state-space methodology seems like a straight line when compared to the wildly varying estimates by the daily model. In the next section it is evaluated whether the relative stability of the state-space model does indeed make its predictions more suitable to use in a real-world context, by using them in a volatility prediction model.

## 4.2 Volatility predictions

Section 4.1 discussed to what degree it is possible to predict the shape of the term-structure curve out-of-sample. We concluded that in general the daily model produces better term-structure predictions. However, it seems to overfit the data, creating predictions that have little error but very extreme estimates of the unobserved level, slope, and curvature factors. At the cost of some predictive accuracy, the estimates of the unobserved factors using the state-space methodology are much more stable and less extreme. This Section evaluates to what extent the predicted slope and curvature factors can be used to improve a baseline volatility model. First, mean absolute errors (MAEs) are presented for all commodities over the full sample for a one-day prediction horizon. Thereafter, results are compared to those using a one-month prediction horizon, and some specific models, commodities, and subsamples are investigated further using graphical measures.

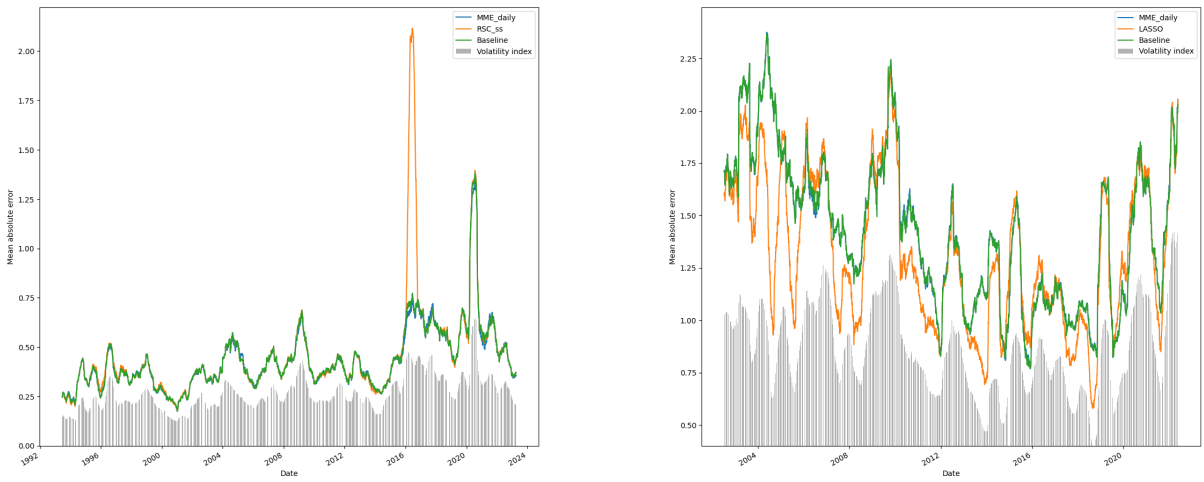
### 4.2.1 Short horizon

Table 4 reports the MAEs as a fraction of the MAE of the baseline HAR model for all other models and all commodities over the whole sample for a one-day prediction horizon. In general,

**Table 4. MAE as a fraction of the baseline model for a one-day prediction horizon.**

| | KW | CL | RB | NG | HO | CT | W | S | SM | BO | C | HG | KC | FC | LC | LH | CO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $ASD^{daily}$ | 1.015 | 1.031 | 1.015 | 1.019 | 1.028 | 1.011 | 1.011 | 1.019 | 1.017 | 1.016 | 1.013 | 1.024 | 1.022 | 1.010 | 1.006 | 1.012 | 1.032 |
| $ACD^{daily}$ | 1.011 | 1.028 | 1.014 | 1.012 | 1.031 | 1.011 | 1.010 | 1.019 | 1.013 | 1.017 | 1.015 | 1.020 | 1.012 | 1.011 | 1.009 | 1.011 | 1.021 |
| $ASD^{ss}$ | 1.007 | 1.042 | 1.046 | 1.055 | 1.014 | 1.046 | 1.036 | 1.016 | 1.045 | 1.012 | 1.109 | 1.038 | 1.031 | 1.010 | 1.018 | 1.019 | 1.028 |
| $ACD^{ss}$ | 1.018 | 1.059 | 1.049 | 1.070 | 1.052 | 1.063 | 1.020 | 1.023 | 1.024 | 1.015 | 1.039 | 1.047 | 1.039 | 1.018 | 1.096 | 1.036 | 1.034 |
| $RSC^{daily}$ | 1.803 | 1.031 | 1.001 | 1.024 | 1.006 | 1.034 | 1.384 | 9.034 | 1.006 | 1.008 | 30.518 | - | - | 1.096 | 1.002 | 1.004 | - |
| $RCC^{daily}$ | 1.898 | 1.012 | 1.002 | 1.019 | 1.019 | 1.034 | 1.366 | 5.917 | 1.006 | 1.008 | 34.806 | - | - | 1.130 | 1.002 | 1.004 | - |
| $RSC^{ss}$ | **0.996**$^{10}$ | 1.029 | 1.018 | **0.998** | 1.006 | 61.936 | 1.006 | 1.010 | 1.016 | 1.027 | 4.651 | 1.049 | 1.008 | 1.054 | 1.019 | 1.005 | 1.031 |
| $RCC^{ss}$ | 1.002 | 1.033 | 1.014 | 1.006 | 1.005 | 46.185 | 1.008 | 1.010 | 1.005 | 1.309 | 1.012 | 1.789 | 1.025 | 1.742 | 1.040 | 1.008 | 1.011 |
| $RSD^{daily}$ | 1.055 | 1.744 | 1.049 | 1.097 | 1.127 | 1.065 | 1.042 | 1.111 | 1.074 | 1.087 | 1.083 | 1.114 | 1.255 | 1.030 | 1.030 | 1.033 | 1.093 |
| $RCD^{daily}$ | 1.082 | 1.305 | 1.069 | 1.037 | 1.174 | 1.088 | 1.060 | 1.102 | 1.108 | 1.087 | 1.083 | 1.062 | 1.210 | 1.087 | 1.109 | 1.036 | 1.113 |
| $RSD^{ss}$ | 1.069 | 1.265 | 1.154 | 1.333 | 1.166 | 1.209 | 1.243 | 1.116 | 1.094 | 1.058 | 1.304 | 1.133 | 1.177 | 1.029 | 1.241 | 1.054 | 1.070 |
| $RCD^{ss}$ | 1.090 | 1.884 | 1.333 | 1.106 | 1.216 | 1.258 | 1.059 | 1.362 | 1.117 | 1.106 | 1.108 | 1.091 | 1.058 | 1.105 | 1.415 | 1.037 | 1.080 |
| $MME^{daily}$ | **0.998** | **0.994**$^{1}$ | **0.994**$^{1}$ | 1.002 | **0.998** | 1.003 | **0.998** | 1.004 | 1.001 | 1.002 | 1.002 | 1.002 | **0.995**$^{1}$ | **0.994**$^{5}$ | 1.003 | **0.999** | 1.002 |
| $MME^{ss}$ | 1.001 | 1.002 | 1.001 | 1.004 | 1.004 | 1.002 | 1.000 | 1.003 | 1.001 | 1.004 | 1.001 | 1.002 | 1.002 | 1.001 | 1.003 | 1.001 | 1.001 |
| $MAE^{daily}$ | 1.002 | 1.001 | 1.001 | 1.007 | 1.001 | 1.000 | 1.002 | 1.000 | 1.000 | 1.009 | 1.002 | 1.002 | 1.003 | 1.000 | 1.005 | 1.000 | 1.002 |
| $MAE^{ss}$ | 1.000 | 1.005 | 1.001 | 1.027 | 1.001 | 1.000 | 1.002 | 1.000 | 1.000 | 1.009 | 1.002 | 1.002 | 1.003 | 1.000 | 1.005 | **0.999** | 1.002 |
| OLS | 2.182 | 1.764 | 1.359 | 1.756 | 1.397 | 93.021 | 2.099 | 34.689 | 1.381 | 1.410 | 85.125 | - | - | 2.299 | 1.541 | 1.228 | - |
| LASSO | 2.753 | 1.077 | 1.065 | **0.987**$^{5}$ | 1.147 | 1.612 | 3.505 | 2.375 | 1.575 | 1.698 | 10.680 | 1.778 | 64.364 | 3.664 | 2.684 | 1.537 | 1.265 |
| NN | - | 1.236 | 1.144 | 1.573 | 1.224 | 99.313 | - | - | 5.026 | 13.937 | - | - | - | 28.609 | 4.211 | 9.542 | - |
| RF | 1.044 | 1.111 | 1.078 | 1.039 | 1.062 | 1.068 | 1.052 | 1.060 | 1.040 | 1.068 | 1.073 | 1.100 | 1.045 | 1.064 | 1.067 | 1.049 | 1.041 |

*Note.* Table reports for each model and commodity the mean absolute error of one-day ahead volatility predictions as a fraction of the error of the baseline HAR model. Numbers in bold indicate commodities for which the model outperformed the baseline. For brevity, the significance of the Diebold-Mariano test is only reported for these situations, with superscript 10 for 10% ,superscript 5 for 5%, and superscript 1 for 1% significance. Moreover, all models that are at least a hundred times worse than the baseline are denoted as -.




**(a) One-day feeder cattle** **(b) One-month natural gas**

**Figure 8. 125-day rolling mean absolute volatility prediction errors**

we notice that it is difficult to beat the baseline model over the full sample, only five models manage to do so, often with a small margin. The models that manage to beat the baseline for at least one commodity are OLS-estimated models based on the relative slope deviation estimated using the state-space methodology, both models based on the curve prediction errors, the one based on the mean absolute daily error and the LASSO including all explanatory variables. Out of these, the directional error model based on the daily estimated factors is the best overall. It is able to beat the baseline model at the 1% significance level for crude oil, gasoline, and coffee.

Figure 8b shows the 125-day rolling MAE for the baseline, the daily error model, and the best non-error-based model. First, we notice that all three models have very similar rolling MAE. All under and outperformance seems to be driven by only a few periods, most clearly characterised by the one large peak in the mean absolute error of the relative slope coefficient model. While the model based on daily errors is the only model able to beat the baseline at the 1% significance level, the LASSO model is actually the one with the smallest error as a fraction of the baseline. This is interesting, since it implies that the error differential series for the LASSO and the baseline model has higher autocovariance than the error differential series for the daily error model and the baseline model. This can be explained at least in part by the fact that the natural gas market is infamously volatile, presumably leading to higher errors, larger differences in errors and thus larger (autoco)variance of the error difference series which is used to calculate the standard Diebold-Mariano test statistic.

An attentive reader might have noticed the remarkably poor performance of some combinations of model and commodity, some having MAE more than a hundred times worse than the baseline HAR model for that commodity. An even more attentive reader would have also

noticed that this only occurs in models that include one or more regressors based on the unbounded slope or curvature deviation. Closer inspection of the errors through time reveals that the inaccuracy of these models is not constant, rather the extremely high MAE can almost fully be attributed to a few periods for each commodity. The peak in rolling MAE shown in Figure 8b is an example of such a period, yet the effect is already smaller here, since the errors are averaged over half a year. As we hypothesised in Section 4.1, these periods can be characterised by an extreme estimate for the slope factor and an equally extreme but opposite in sign estimate for either the level or curvature factor. Combined with high decay, these estimates lead to a relatively normal prediction of price curves such as the one shown in 4b. However, the volatility model knows only the estimated slope and/or curvature factors and thus 'thinks' that the market is experiencing an extreme divergence from the regular term structure, leading to extreme under- or overprediction of volatility. The daily model estimates the parameters such that they fit a single day best, while the state-space model also has to estimate the relations between days. Therefore, the daily model is more likely to produce these extreme estimates of the slope and curvature factors, explaining why the problem is even larger for the versions of the model that use the daily estimated slope and curvature factors.

### 4.2.2 Long horizon

To investigate whether the slope and curvature factors might contain longer-term volatility signals, the models are also used to create one-month-ahead predictions. The results are reported in Table 5. Similarly to the one-day prediction horizon, the models that can beat the baseline are those based on mean (absolute) errors. The relative slope coefficient model based on the state-space estimation was able to beat the baseline on the short horizon. However, it seems to have lost this predictive power on the longer horizon. The LASSO model, on the other hand, is now the best for crude oil and natural gas, even managing to reduce the mean absolute error of the baseline model for natural gas by almost 7%.

To investigate where the LASSO model gets its advantage, we compare it graphically against the baseline HAR and the second best model for the natural gas market, the $MME^{daily}$ model. The predictions of the baseline HAR model and the one with the added variable $MME^{daily}$ are so similar that Figure 8a seems to contain only two lines, rather than three. The figure shows the half-year rolling mean absolute error for the volatility predictions. It seems as though the $MME^{daily}$ model produces slightly better volatility predictions between 2005 and 2006, and slightly worse ones in the period 2010 to 2011. On the contrary, the predictions made by the LASSO model are clearly distinguishable from those by the baseline model. Moreover, the

**Table 5. MAE as a fraction of the baseline model for a one-month prediction horizon.**

| | KW | CL | RB | NG | HO | CT | W | S | SM | BO | C | HG | KC | FC | LC | LH | CO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ASD$^{daily}$ | 1.010 | 1.021 | 1.024 | 1.021 | 1.010 | 1.012 | 1.019 | 1.035 | 1.020 | 1.018 | 1.021 | 1.018 | 1.011 | 1.017 | 1.009 | 1.009 | 1.016 |
| ACD$^{daily}$ | 1.014 | 1.072 | 1.022 | 1.038 | 1.026 | 1.011 | 1.015 | 1.031 | 1.014 | 1.015 | 1.019 | 1.017 | 1.006 | 1.017 | 1.014 | 1.009 | 1.019 |
| ASD$^{ss}$ | 1.031 | 1.161 | 1.104 | 1.054 | 1.021 | 1.046 | 1.025 | 1.039 | 1.052 | 1.026 | 1.053 | 1.080 | 1.068 | 1.032 | 1.044 | 1.022 | 1.048 |
| ACD$^{ss}$ | 1.038 | 1.121 | 1.115 | 1.079 | 1.030 | 1.037 | 1.030 | 1.056 | 1.032 | 1.033 | 1.052 | 1.114 | 1.069 | 1.054 | 1.053 | 1.026 | 1.026 |
| RSC$^{daily}$ | 1.927 | 1.030 | 1.001 | 1.135 | 1.006 | 1.059 | 2.881 | 21.567 | 1.036 | 1.003 | 88.936 | - | - | 1.112 | 1.003 | 1.001 | - |
| RCC$^{daily}$ | 1.876 | 1.017 | 1.004 | 1.059 | 1.021 | 1.059 | 2.874 | 6.358 | 1.037 | 1.003 | 19.441 | - | 99.782 | 1.069 | 1.003 | 1.001 | - |
| RSC$^{ss}$ | 1.014 | 1.044 | 1.143 | 1.051 | 1.030 | - | 1.045 | 1.046 | 1.080 | 1.106 | 3.911 | 1.060 | 1.038 | 1.213 | 1.086 | 1.035 | 1.121 |
| RCC$^{ss}$ | 1.037 | 1.063 | 1.116 | 1.061 | 1.038 | 39.807 | 1.053 | 1.047 | 1.019 | 2.484 | 1.040 | 1.411 | 1.060 | 1.600 | 1.059 | 1.027 | 1.031 |
| RSD$^{daily}$ | 1.037 | 1.079 | 1.052 | 1.034 | 1.075 | 1.034 | 1.063 | 1.100 | 1.150 | 1.062 | 1.064 | 1.067 | 1.046 | 1.034 | 1.010 | 1.037 | 1.016 |
| RCD$^{daily}$ | 1.039 | 1.089 | 1.046 | 1.117 | 1.047 | 1.043 | 1.063 | 1.123 | 1.061 | 1.055 | 1.087 | 1.044 | 1.048 | 1.038 | 1.094 | 1.062 | 1.035 |
| RSD$^{ss}$ | 1.109 | 1.550 | 1.306 | 1.233 | 1.088 | 1.138 | 1.081 | 1.085 | 1.091 | 1.082 | 1.035 | 1.129 | 1.080 | 1.034 | 1.180 | 1.067 | 1.304 |
| RCD$^{ss}$ | 1.053 | 1.846 | 1.316 | 1.142 | 1.071 | 1.214 | 1.122 | 1.130 | 1.095 | 1.105 | 1.052 | 1.230 | 1.081 | 1.094 | 1.073 | 1.007 | 1.058 |
| MME$^{daily}$ | 1.000 | **0.998** | 1.000 | 1.002 | **0.999** | 1.002 | 1.002 | 1.002 | 1.001 | 1.000 | 1.002 | 1.002 | 1.002 | **0.998**$^{10}$ | 1.000 | 1.000 | 1.000 |
| MME$^{ss}$ | 1.001 | 1.003 | 1.002 | 1.002 | 1.003 | 1.001 | 1.002 | 1.002 | 1.002 | 1.001 | 1.002 | 1.001 | 1.003 | 1.001 | 1.000 | 1.002 | 1.001 |
| MAE$^{daily}$ | 1.002 | 1.010 | **0.999** | 1.012 | 1.002 | 1.003 | 1.003 | **0.999** | 1.000 | 1.001 | 1.000 | 1.001 | 1.003 | 1.000 | 1.005 | 1.000 | 1.002 |
| MAE$^{ss}$ | 1.002 | 1.015 | 1.001 | 1.071 | 1.004 | 1.009 | 1.003 | 1.000 | 1.002 | 1.006 | 1.008 | 1.002 | 1.002 | 1.000 | 1.005 | 1.000 | 1.002 |
| OLS | 2.052 | 1.703 | 1.321 | 1.973 | 1.266 | - | 3.788 | 10.325 | 1.482 | 2.005 | 21.405 | - | - | 2.014 | 1.383 | 1.241 | - |
| LASSO | 3.952 | **0.984** | 1.002 | **0.930**$^{1}$ | 1.091 | 1.575 | 2.919 | 2.096 | 1.517 | 1.582 | 4.540 | 1.815 | 1.100 | 3.389 | 2.581 | 1.462 | 1.235 |
| NN | - | 1.287 | 1.072 | 1.894 | 1.118 | - | - | 5.027 | 1.080 | 17.880 | - | - | - | 59.620 | 15.371 | 6.997 | - |
| RF | 1.072 | 1.086 | 1.052 | 1.109 | 1.105 | 1.066 | 1.101 | 1.091 | 1.080 | 1.083 | 1.101 | 1.141 | 1.107 | 1.095 | 1.112 | 1.080 | 1.058 |

*Note.* Table reports for each model and commodity the mean absolute error of one-month ahead volatility predictions as a fraction of the error of the baseline HAR model. Numbers in bold indicate commodities for which the model outperformed the baseline. For brevity, the significance of the Diebold-Mariano test is only reported for these situations, with superscript 10 for 10%, superscript 5 for 5%, and superscript 1 for 1% significance. Moreover, all models that are at least a hundred times worse than the baseline are denoted as -.

LASSO model is almost exclusively equal or better than the baseline, regardless of the specific half-year period considered. Figure 8a also shows an indexed version of the 125-day rolling mean volatility, given by the grey bars. While the baseline and LASSO models perform similarly in high volatility periods, the LASSO model is able to achieve a relatively higher error reduction in the low-volatility regimes. Much of this seems to be driven by it adapting much faster to a decrease in measured volatility than the baseline model.

For both the short and long prediction horizons, the regularised LASSO regression is the only ML method to outperform the baseline model. Moreover, the Neural Network is one of the worst performing methods overall, often having a MAE that is more than a hundred times worse than the baseline model. At first this result seems surprising, since neural networks have been put to great use recently in a broad spectrum of research directions. However, when we consider that a neural network is also a combination of linear effects, it makes sense that very extreme estimates of the slope and curvature factors can lead to short periods with very high prediction error, much like is this case with regular OLS. The issue is amplified since the neural network contains both the daily and state-space estimated factors as explanatory variables, and only one of these needs to have an extreme estimate to disrupt the prediction of the neural network. Lastly, the basic structure of the one-layer perceptron is the type of neural net most vulnerable to these problems, as the more complex neural networks can 'ignore' the extreme effect through extra layers and activation functions.

# 5 Conclusion

Commodities are an asset class very different from stocks and bonds. The price of each commodity is influenced by various supply and demand factors that are often seasonal in nature. Furthermore, trading is done through futures contracts, which introduces a term structure to commodity markets. This paper has investigated to what extent the physical characteristics of commodities are intertwined with the observed term structure curve given by the prices of their futures contracts. It did so by first evaluating the predictability of the shape of the curve using an adapted form of the Nelson-Siegel model. The predictions for the shape were then used as extra explanatory variables in a basic volatility model to investigate to what extent the shape of the curve holds predictive information about volatility in the underlying commodity market.

To predict the shape of the term structure curve out-of-sample, we estimated the adapted Nelson-Siegel model both using a state-space approach and a more parsimonious daily version. The daily model is better able to predict the price curve on a one-day prediction horizon for all commodities. However, the difference in performance is much larger for some commodities than for others. We found that the state-space model for some commodities is unable to fit a curve nicely through the observed prices, and therefore underpredicts the variation caused by the time-to-maturity effect and overpredicts the effects of the seasonal patterns to compensate. Although the daily-estimated model mostly performs well, it is prone to overfitting in order to predict the observed prices correctly, rather than the full price curve. We concluded that the daily model is better able to predict prices, but the state-space model might provide a more accurate prediction of the shape of the full curve.

Following this result, the predictions of the slope and curvature of the price curve made by both models and the prediction errors made by these models were used as additional explanatory variables in a basic HAR volatility model. In general, the baseline model is hard to beat for most commodities. Considering the short one-day prediction horizon, the only standard models able to beat the baseline were the one based on the difference in predicted slope compared to average slope estimated using the state-space methodology and the one based on the prediction errors made by the daily model. Additionally, a LASSO model containing all possible predictors based on the predicted shape of the curve is able to significantly outperform the baseline model for one market at the short term prediction horizon. For the longer one-month prediction horizon, the baseline model proved even harder to beat. However, the LASSO model is still able to generate significantly better volatility predictions for the natural gas market. It was shown that it does so by being consistently about equal or better than the baseline model, since the information implied in the shape of the term structure allows it to adapt faster to low-volatility periods.

The dataset used in this paper is the broadest to date, featuring over 50 years of data for 19 commodities. Therefore, the results presented in this article can be used as an excellent starting point for further academic research. Specific time periods, commodities, or relations between commodities can be examined more in-depth based on the reported results here. For example, one could wonder why the number of actively traded contracts over time varies so much for some commodities such as gold and natural gas, while it seems very stable for agricultural commodities. Furthermore, this paper has shown that there is potential to predict more and less volatile periods using the shape of the term structure curve. This finding can be used by businesses to reduce risks in both the purchasing and sales departments. In addition, governments can make sure essential sectors are prepared for price swings.

However, as this paper is the first to apply Nelson-Siegel type models to such a large dataset with various methods of estimation methods, there exist problems that can be removed in further research or real-world implementations. Most importantly, research should be done to estimate the models using fixed decay, rather than treating decay as a parameter to be optimised. Because of the non-linear effect of decay, it can be tricky to optimise, increasing the likelihood of other estimation issues, such as the extreme estimates produced by the daily model or the flat curve of the state-space model. Fixing the decay limits the total variety of shapes the estimated curve can assume. However, we expect it to be limited to more natural shapes that are more likely to occur in financial markets. Additionally, it could be interesting to find a way to have the strength of the seasonal effect evolve with the price of the commodity, or have it be in some other way time-varying to accommodate changing market dynamics.

## References


Bianchi, R. J., Fan, J. H., Miffre, J., & Zhang, T. (2023). Exploiting the dynamics of commodity futures curves. *Journal of Banking & Finance*, *154*, 106965.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Brennan, M. J. (1976). The supply of storage. *The economics of futures trading*, 100–107.

Christensen, J. H., Diebold, F. X., & Rudebusch, G. D. (2011). The affine arbitrage-free class of nelson–siegel term structure models. *Journal of Econometrics*, *164*(1), 4–20.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*(2), 174–196.

Deaton, A., & Laroque, G. (1992). On the behaviour of commodity prices. *The Review of Economic Studies*, *59*(1), 1–23.

Deaton, A., & Laroque, G. (1996). Competitive storage and commodity price dynamics. *Journal of Political Economy*, *104*(5), 896–923.

Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of econometrics*, *130*(2), 337–364.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134–144.

Diebold, F. X., Rudebusch, G. D., & Aruoba, S. B. (2006). The macroeconomy and the yield curve: A dynamic latent factor approach. *Journal of econometrics*, *131*(1-2), 309–338.

Gorton, G. B., Hayashi, F., & Rouwenhorst, K. G. (2013). The fundamentals of commodity futures returns. *Review of Finance*, *17*(1), 35–105.

Grønborg, N. S., & Lunde, A. (2016). Analyzing oil futures with a dynamic nelson-siegel model. *Journal of Futures Markets*, *36*(2), 153–173.

Heidorn, T., Mokinski, F., Rühl, C., & Schmaltz, C. (2015). The impact of fundamental and financial traders on the term structure of oil. *Energy Economics*, *48*, 276–287.

Hicks, J. R. (1939). The foundations of welfare economics. *The economic journal*, *49*(196), 696–712.

Hirshleifer, D. (1989). Determinants of hedging and risk premia in commodity futures markets. *Journal of Financial and Quantitative Analysis*, *24*(3), 313–331.

Hirshleifer, D. (1990). Hedging pressure and futures price movements in a general equilibrium model. *Econometrica*, 411–428.

Kaldor, N. (1939). Speculation and economic stability. *The review of economic studies*, *7*(1), 1–27.

Karstanje, D., Van Der Wel, M., & van Dijk, D. J. (2017). Common factors in commodity futures curves. *Available at SSRN 2558014*.

Keynes, J. M. (1923). Some aspects of commodity markets. *Manchester Guardian Commercial: European Reconstruction Series*, *13*, 784–786.

Mardia, K. V., Jupp, P. E., & Mardia, K. (2000). *Directional statistics* (Vol. 2). Wiley Online Library.

Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of Business*, 473–489.

Pindyck, R. S. (1990). Inventories and the short-run dynamics of commodity prices.

Qu, R., Timmermann, A., & Zhu, Y. (2023). Comparing forecasting performance with panel data. *International Journal of Forecasting*.

Routledge, B. R., Seppi, D. J., & Spatt, C. S. (2000). Equilibrium forward curves for commodities. *The Journal of Finance*, *55*(3), 1297–1338.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

West, J. (2012). Long-dated agricultural futures price estimates using the seasonal nelson-siegel model. *International Journal of Business and Management*, *7*(3), 78–93.

Working, H. (1949). The theory of price of storage. *The American economic review*, *39*(6), 1254–1262.

# Appendix

## A Rolling procedure

This appendix describes the exact procedure used to roll futures contracts. Figure A shows a timeline relating the Last Trading Date (LTD) to the First Notice Date (FND) for each commodity. Furthermore, any commodities with their mnemonic above the timeline are financially settled, while those below are physically settled. With the exception of FC, the commodities can be placed neatly into three brackets based on the difference in days between the LTD and FND. For each of these brackets we roll all contracts seven days before the first FND in the bracket. FC forms its own bracket and is also rolled seven days before its FND. Finally, Table A gives an overview of the roll dates for each commodity.

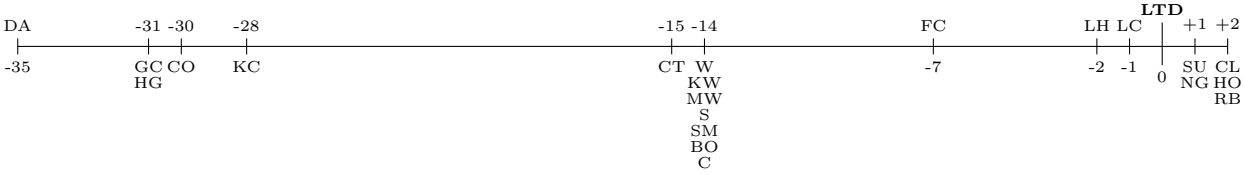



**Figure A 1. Difference in days between the first notice date and last trading date.**

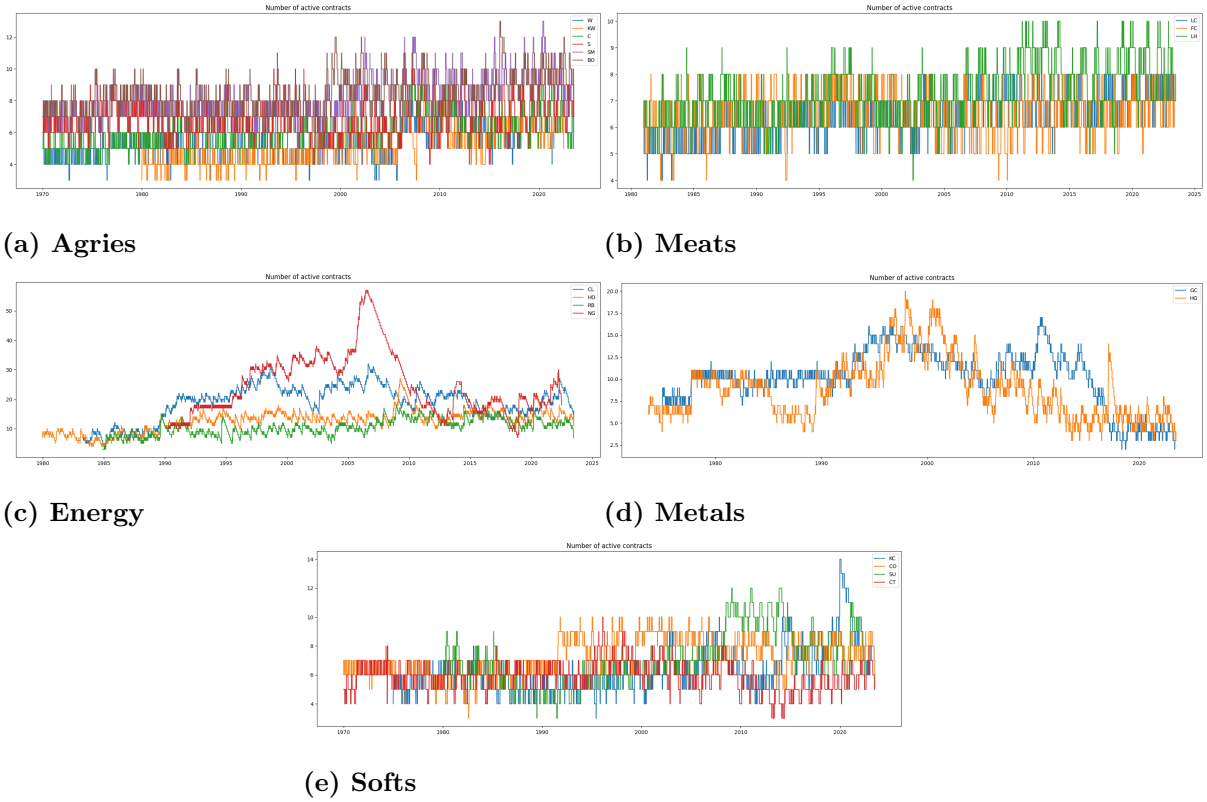
**Table A 1. Rolling dates as a function of their last trading date**

| Bracket | Commodities | Rolled on |
|---|---|---|
| Short | LH, LC, SU, NG, CL, RB, HO | 9 days before their last trading date |
| Medium | CT, W, KW, MW, S, SM, BO, C | 22 days before their last trading date |
| Long | DA, GC, HG, CO, KC | 42 days before their last trading date |
| Feeder Cattle | FC | 14 days before its last trading date |

*Note.* Table reports for each basket which commodities it contains and how long before their last trading date these commodities are rolled over.

# B Active contracts

This Appendix provides the number of contracts defined as active throughout the sample. The commodities are order by their sectors.




**(a) Agries** **(b) Meats**

**(c) Energy** **(d) Metals**

**(e) Softs**

**Figure A 2. Number of active contracts throughout the sample**

# C Hyperparameters

This Appendix provides the relevant hyperparameters used for the LASSO regression, neural network, and the random forest. These are supposed to be standard values and are not optimised as the goal of this research is not to produce the best performing model, but rather to investigate the possibility of using machine learning in this novel context.

**Table A 2. LASSO Hyperparameters**

| **Hyperparameter** | **Value** |
|---|---|
| method | 'LASSO' |
| alpha | 1.0 |
| fit_intercept | True |
| start_params | None |
| profile_scale | False |
| refit | False |

**Table A 3. Neural Network Hyperparameters**

| **Hyperparameter** | **Value** |
|---|---|
| Activation | relu |
| Alpha | 0.0001 |
| Batch size | auto |
| Beta 1 | 0.9 |
| Beta 2 | 0.999 |
| Early stopping | False |
| Epsilon | $1 \times 10^{-8}$ |
| Hidden layer sizes | (100,) |
| Learning rate | constant |
| Learning rate init | 0.001 |
| Max fun | 15000 |
| Max iter | 1000 |
| Momentum | 0.9 |
| N iter no change | 10 |
| Nesterovs momentum | True |

Continued on next page

Table A 3 – continued from previous page

| Hyperparameter | Value |
|---|---|
| Power t | 0.5 |
| Random state | None |
| Shuffle | True |
| Solver | adam |
| Tol | 0.0001 |
| Validation fraction | 0.1 |
| Verbose | False |
| Warm start | False |

**Table A 4. Random Forest Hyperparameters**

| Hyperparameter | Value |
|---|---|
| n_estimators | 100 |
| criterion | 'squared_error' |
| max_depth | None |
| min_samples_split | 2 |
| min_samples_leaf | 1 |
| min_weight_fraction_leaf | 0.0 |
| max_features | 1.0 |
| max_leaf_nodes | None |
| min_impurity_decrease | 0.0 |
| bootstrap | True |
| oob_score | False |
| n_jobs | None |
| random_state | None |
| verbose | 0 |
| warm_start | False |
| ccp_alpha | 0.0 |
| max_samples | None |
| monotonic_cst | None |