Evaluating Oil Price Forecasts: A meta-analysis

Michail Filippidis* George Filis† Georgios Magkonis‡

February 24, 2022

Abstract

Oil price forecasts have traditionally attracted the interest of both the empirical literature and policy makers, although research efforts have intensified in the last 15 years. The present study investigates which are the forecasting characteristics that have the greatest impact on the accuracy level of such forecasts. To do so, we collect more that 6,000 observations of root mean squared errors and we show that combined forecasts as well as MIDAS models tend to report significantly lower forecast errors. In addition, the choice of the oil price benchmark is also important, with the Brent crude oil prices to be better predicted. Also, the forecasting horizon proves to be an important factor and the same holds for real, instead of nominal, prices. A number of robustness tests confirm the validity of these results. The findings of this study serve as a guide for oil price forecasting exercises.

Keywords: Oil price forecasts, meta-analysis, model averaging

JEL Codes: Q47, C83, C53

*Economics & Finance Subject Group, Portsmouth Business School, University of Portsmouth, Portland street, Portsmouth, PO1 3DE, UK

†Department of Economics, University of Patras, University Campus, 265 04, Rio, Patras, Greece

‡Economics & Finance Subject Group, Portsmouth Business School, University of Portsmouth, Portland street, Portsmouth, PO1 3DE, UK

1 Introduction

Since the regime change in oil price fluctuations in the early 2000, which is characterised by the unprecedented levels of oil prices and severe volatility, the extant literature (Knetsch, 2007; Alquist et al., 2013; Degiannakis and Filis, 2018) and policy documents (Bernanke, 2005; ECB, 2015) have highlighted the ever more important need for accurate oil price forecasts. Figure 1 depicts this regime change in oil prices in the post-2003 period, when we observe a series of episodes with huge price swings. For instance, the WTI crude oil price reached its peak at almost 140 dollars in July 2008, which was then followed by a rapid and sharp decline at about 40 dollars in January 2009. In addition, in June 2014, oil reached once again a price above 100 dollars, and then a fall at below 50 dollars seven months later (February 2015). More recently, due to Covid-19 pandemic, oil prices lost about 65% of their value (from 60 dollars in December 2019 to almost 20 dollars in April 2020). We should also highlight that for the first time we experienced negative oil prices, when the WTI dropped at -37 dollars on the 20th April 2020 (not shown in Figure 1 as it was constructed using monthly data).

Figure 1 here

The need for accurate oil price forecasts, given the aforementioned abrupt changes, stems from the fact that they form important decision-making inputs for a number of stakeholders, including private businesses, central banks and the national governments. For instance, Alquist et al. (2013) provide evidence that oil price forecasts help industrial sector companies to forecast their product prices. Moreover, they indicate that investment decisions regarding climate change and carbon emissions predictions, as well as, formulation of regulatory policies in the energy sector may significantly be influenced by oil price forecasts. In addition, Baumeister (2014) shows that oil price forecasting is an important tool for monetary authorities given that it conveys information about predictions in inflation and economic activity. Finally, Baumeister et al. (2018) highlights the importance of oil price forecasts for national

governments of both oil-exporters and oil-importers in devising their investment strategies and budget plans. It is also important to highlight that central banks are interested in forecasting real oil prices in domestic currency units, which captures the real cost of oil for domestic consumption (Baumeister and Kilian, 2014). In turn, this further increases the complexity of oil price forecasts as their accuracy further depends of how future exchange rates are estimated.

Moreover, since oil is a physical commodity, it is intuitively expected that its price should be primarily affected by oil market fundamentals, namely unexpected oil supply disruptions, unanticipated changes in global demand for crude oil and unexpected changes in inventory demand (see, for instance, Kilian (2009), Kilian (2010), Kilian and Murphy (2014)). However, the more recent literature highlights the significant effect of financial markets as drivers of oil price movements. This is known as the financialisation of the oil market and is primarily related to speculative activity in this market. In this regard, Fratzscher et al. (2014) explain that oil acts as a financial asset due to the fact that it reacts rapidly to information associated with other financial assets such as stock prices or exchange rates. In this regard, Degiannakis and Filis (2018) show that apart from the oil market fundamentals, information stemming from the financial markets could improve oil price forecasts.

Based the aforementioned developments in the oil market, as well as, the complexity of its price forecasts, the literature has developed an array of different modelling frameworks and has employed a series of different predictors in search of improved accuracy. For instance, early studies focus on the use of Vector Error Correction models (Coppola, 2008; Murat and Tokat, 2009) or futures-based forecasts (Knetsch, 2007; Alquist and Kilian, 2010) and try to show whether these forecasts can outperform the random-walk. Other studies, such as those by Baumeister and Kilian (2012), Baumeister and Kilian (2014) and Naser (2016) employ Vector Autoregressive-type (VAR) models (e.g. structural VARs, time-varying parameter VARs), whereas Baumeister et al. (2014) and Baumeister and Kilian (2014, 2015) assess the predictive accuracy of forecasting averaging. Baumeister et al. (2015) and, more recently,

Degiannakis and Filis (2018) exploit the advantages of the mixed-data sampling (MIDAS) model.

In terms of predictors, the aforementioned studies have more commonly used the oil market fundamentals, although it is not uncommon for studies to use product spreads or futures prices. Typically, the data are obtained in monthly frequency. Nevertheless, the recent studies by Baumeister et al. (2015) and Degiannakis and Filis (2018) assess whether financial data that can be obtained in higher frequency (daily or weekly) improve the forecasting accuracy of oil prices. The choice of the crude oil price benchmark is another interesting distinction among the aforementioned studies. There are three main variables, namely, the US refiners' acquisition cost (RAC), the West Texas Intermediate (WTI) and the Brent crude oil (Brent), which are used in the forecasting exercises.

Given the increased interest in this line of research and the numerous papers that have been published, especially in the last 15 years, it is timely to proceed with a quantitative navigation that will allow us to explore whether there are factors that systematically provide better oil price forecasts. To achieve this, we employ a meta-analysis approach, which has been proven to be a useful tool for summarising the empirical findings of existing research.[1] To do so, we measure the impact of the different forecasted variables, model classes, horizons, as well as, time periods, on the forecast error. Overall, the contribution is to identify the factors that play an important role in oil price prediction.

In short, our empirical findings can be described as follows. First, MIDAS and combined forecasting models, among other models, exhibit significantly higher forecasting accuracy. Second, the use of Brent crude oil prices generates better predictions in comparison with other crude oil benchmarks. Finally, the short-run forecasting horizons and the use of real oil price also contribute to lower forecast errors.

The remaining of the paper is structured as follows. Section 2 describes the data collection process. Section 3 presents the methodology, while Section 4 discusses the main results along

[1]Meta-analytic techniques have been extensively employed in the energy economics discipline; see for instance Havranek and Kokes (2015), Mattmann et al. (2016) and Dimitropoulos et al. (2018).

with robustness checks. Finally, Section 5 concludes the study.

2 Forecasts errors across literature

2.1 Data collection

A significant part of the forecasting literature uses the random walk (RW) model without a drift, or the no-change forecast, as the benchmark model. Its h-step ahead forecast error at any time point is shown as $e^{RW}_{t+h|t} = o_{t+h|t} - o_{t+h}$, where o represents the price of oil. Furthermore, the root mean squared error (RMSE, thereafter) is used as the main loss function to measure the forecasting accuracy. The forecasting accuracy of the RW model would be as follows:

$$RMSE^{RW} = \sqrt{\sum_{t=i}^{T} (e^{RW}_{t+h|t})^2 / T}, \tag{1}$$

where $t = i...T$ denotes the out-of-sample forecasting observations. It is also very common for studies to report the forecasting performance of competing models in relative terms against the RMSE of the RW model, such as that:

$$RRMSE^m_{kl} = \psi^m_{kl} = (RMSE^m_{kl})/(RMSE^{RW}_{kl}), \tag{2}$$

where the superscript m denotes a forecasting model other than the random walk, while k and l indicate the forecast error k from the study l. When the ratios are lower than one then the competing model is able to outperform the benchmark model. The main variable of interest of our study is the relative RMSE ($RRMSE$, thereafter).[2] We use the terms $RRMSE^m_{kl}$, ψ^m_{kl} and 'relative forecasting performance' interchangeably throughout the text.

We perform a Google scholar search using the key combinations 'oil price predictability', 'oil price forecasts', 'oil price forecasting' and 'oil price modelling'. In order to impose a

[2]See Eickmeier and Ziegler (2008).

certain quality threshold, we only focus on published papers. The inclusion of working papers, many of which produce large horse-races of forecasts, would make the total amount of observations intractable. This search was carried out between July and September 2021. The next step is to impose a number of criteria according to which a study can be included in our sample. Our first criterion requires for a study to report at least one RMSE, which is the metric that we focus on. Therefore, the papers that report only alternative evaluation metrics for the relative forecasting performance (such as the direction of change statistic) are excluded. The second criterion is associated with the great number of combinations of models and the corresponding RRMSE, and thus, we collect only on the reported RMSE that use random walk as the benchmark model.[3] So, the reported RMSEs or RRMSEs that are based on different benchmark are not included in our meta-sample. Our third criterion is related to the use of non-standard machine learning models (such as random forests and artificial neural networks) from a number of papers the last years. As a result, we focus on traditional econometric models and thus we excluded those studies. In this way, we ensure the comparability of the collected RRMSEs across studies. Overall, the total sample consists of 6,089 observations collected from 21 papers. Table 1 presents descriptive statistics across these selected studies. Figure 2 presents the histogram of the selected RRMSEs, which exhibit fat tails, showing a wide range of forecast errors. The selection process is summarised in a PRISMA chart in Appendix 1, while the full list of studies is provided in Appendix 2.

Table 1 here

Figure 2 here

[3]Benmoussa et al. (2020) argue that the conventional random walk forecast is uninformative in terms of forecast accuracy and should not be used for forecast comparisons for aggregated data. However, our decision to employ the random walk based on the fact that this benchmark is widely used in the literature on forecasting oil prices.

2.2 Heterogeneity of forecasts errors

In order to examine the heterogeneity of the reported RMSEs across the literature, we look into three categories. From each category we identify factors that may systematically influence the reported RMSEs. In what follows we describe this process.

Oil price. Primarily, we focus our attention on the oil price that is forecasted in the published studies. We posit that there could be oil price benchmarks that are harder to predict. If this holds true, forecasters should be aware so as to engage in efforts to improve their forecasting frameworks for these benchmarks. We discern three major choices among the collected papers. The first is the West Texas Intermediate (WTI). Therefore, we create a dummy variable that takes 1 when the RRMSE comes from a forecasting exercise that uses WTI and 0 otherwise. Secondly, we consider the US refiners' acquisition cost (RAC) as an alternative price index. In a similar vein, we use a dummy variable that takes 1 when the RRMSE comes from an analysis that uses RAC and 0 otherwise. The final choice is the use of Brent. We treat Brent as the base category. Another feature of the oil price is whether the forecasting exercise uses the nominal or real price. So, we create an additional moderator variable ('real') that takes 1 for real and 0 for nominal oil price.

Forecasting Models. The choice of forecasting models is the most challenging one. Firstly, we treat as the base category the relative forecasting performance of the models that belong to the ARIMA family, i.e.,

$$\psi_{kl}^{ARIMA} = (RMSE_{kl}^{ARIMA})/(RMSE_{kl}^{RW}). \tag{3}$$

Secondly, we create the moderator variable 'structural' that takes 1 when the RRMSE comes from a structural model, either a structural VAR or a DSGE. Thirdly, we create the dummy variable 'midas' where a value of 1 is assigned when a MIDAS model is used. In a similar fashion, we use the same type of moderator dummy variables for models using regression-based forecasts ('regression'), combined forecasts ('combined') and finally, models

that use futures prices ('future') and product spreads ('product').[4]

Forecasting Features. Finally, we consider two features of the forecasting exercise. The first characteristic is the forecasting horizon. The collected studies that constitute our meta-sample use different data frequencies. Therefore, the forecasting horizon is expressed in different frequencies. To overcome this problem, we convert the horizons into months in order to obtain a homogeneous measure across all studies. The second characteristic is the forecasting period. More precisely, we take into account the date of the end-of-sample, as the forecast errors are reported for the end of the sample period. In this way, we are able to test whether there is a trend in the reported results. Since Baumeister and Kilian (2012) the real-time forecasts are standard in the literature. We additionally take this characteristic into account by including the variable 'realtime' that takes 1 when real-time forecasts are reported.

Table 2 presents the moderator variables analysed above, while Figure 3 provides a graphical illustration of the heterogeneity of the reported estimates across the three different categories. For the case of forecasting horizons, we follow Chinn and Meese (1995) and we consider as short-run the horizons up to 12 months and as long-run the ones that are more than 12 months.

Table 2 here

Figure 3 here

3 Methodology

This section presents the method according to which the factors that systematically affect the reported estimates can be identified. The benchmark method is the Bayesian model averaging (BMA) that belongs to the family of models that deal with big data (Koop, 2017). The usefulness of this technique is properly revealed when the number of regressors is quite

[4]See Degiannakis and Filis (2018) for a brief discussion of these methods.

large. Overall, BMA remains an increasingly popular method of identifying the significant drivers of a specific variable (here the RRMSE). Our meta-regression model can be written as:

$$\psi_{kl}^{m} = c + \sum_{S=1}^{12} \gamma_S^{\xi} Z_{S,kl} + \epsilon_{kl}, \tag{4}$$

where ψ_{kl}^{m} is the k relative forecast error from the l study for forecasting model m, the Z matrix contains the moderator variables described in Section 2.2, γ_s are the coefficients of each moderator S, while ϵ is the error term that is normally distributed as $N(0, \sigma)$. The superscript ξ indicates that the equation (4) is valid under model M_ξ of the BMA exercise. In our case, the use of 12 regressors results to 4,096(=2^{12}) different models to choose from. This means that the model space consists of M_1,...,M_ξ models, where $\xi \in [1, ..., 4096]$. Due to the moderate amount of explanatory variables, it is computationally feasible to evaluate and average over all the models. At the same time, the analytical solution can also be derived. We estimate the posterior model density as well as the posterior inclusion probabilities both analytically and computationally. The results from both approaches are identical.

The main characteristic of the model averaging techniques is that they assign a weight to each model and then, average across these models. So, the inference is not based on individual models, but on weighted averages. Even with a small number of regressors, the model space consists of many potential combinations. In the remaining part of this Section, we present the basic concepts of the BMA. Appendix 3 provides a more detailed technical discussion. Based on the Bayes' rule, the posterior density of γ is written as:

$$p(\gamma|\psi, Z) = \sum_{j=1}^{4096} p(\gamma_j|\psi, Z, M_j) p(M_j|\psi, Z), \tag{5}$$

where $p(\gamma_j|\psi, Z, M_j)$ is the posterior distribution under model M_j and $p(M_j|\psi, Z)$ is the posterior model probability.[5] The above equation shows that the posterior model probabilities

[5]To avoid unnecessary confusion, we will use ψ instead of ψ_{kl}^{m} for the remaining of the paper.

are used as weights. More precisely, the posterior density of γ for each model M_j is weighted by the posterior model probability of each model M_j. The point estimates for the posterior mean can be derived by taking expectations:

$$E(\gamma|\psi, Z) = \sum_{j=1}^{4096} E(\gamma_j|\psi, Z, M_j)p(M_j|\psi, Z). \tag{6}$$

The posterior variance is proved to be:

$$\begin{aligned} Var(\gamma|\psi, Z) = \sum_{j=1}^{4096} p(M_j|\psi, Z)Var(\gamma_j|\psi, Z, M_j) + \\ \sum_{i=j}^{4096} p(M_j|\psi, Z)(E(\gamma_j|\psi, Z, M_j) - E(\gamma_j|\psi, Z, M_j)^2). \end{aligned} \tag{7}$$

The most useful metric is the posterior inclusion probability (PIP) which is defined as the sum of posterior model probabilities of all the models that include the specific regressor and takes the following form:

$$PIP_i = \sum_{j=1} p(M_j|\psi, Z), \tag{8}$$

with $i \in [1, 12]$ indicating that each regressor has a specific inclusion probability. Therefore, the PIP shows how frequently a regressor appears in the alternative M_j models. In this way, the level of PIP determines whether a regressor can be considered as a robust determinant. The closer to 1 the PIP of a regressor is, the larger its explanatory power. In other words, the variable with the highest estimated PIP is this variable that is present in almost all the alternative models and therefore, a robust driver that explains the heterogeneity of the reported estimates.

As far as the parameters priors are concerned, we choose the following options. As there is no prior knowledge, we use non-informative priors for the intercept and the variance; $p(c) \propto 1$ and $p(\sigma) \propto \sigma^{-1}$. Regarding the γ parameters, we assume that they are centered at zero and

the variance is proportional to $\sigma^2(g(Z_iZ_i)^{-1})$, where g is the Zellner's g hyperparameter that indicates the level of uncertainty. The larger the g is, the larger the prior coefficient variance and, therefore, the higher our uncertainty. In summary, the coefficients' distribution depends on g:

$$\gamma_i|g \sim N(0, \sigma^2(g(Z_iZ_i)^{-1}). \tag{9}$$

In this study, we employ two different choices regarding g. Firstly, we set $g = N$, which is the unit information prior (UIP), where N is the sample size. Secondly, we set the hyper-g prior as suggested by Liang et al. (2008). For the case of model priors, we also use two alternative choices. Firstly, we use the uniform model prior that assumes equal probability to all models. Secondly, we relax this assumption by setting a beta-binomial prior. The approximation of the posterior distribution is simulated by a MCMC sampler algorithm.

4 Results

4.1 Main evidence

Table 3 shows the first round of results. Following Kass and Raftery (1995), we categorise the effect of a variable as weak, positive, strong, and decisive if its PIP lies between 0.5-0.75, 0.75-0.95, 0.95-0.99 and 0.99-1, respectively. We begin our analysis with the evaluation of the relative performance of the different forecasting frameworks. The results suggest that the MIDAS models, as well as, the combined forecasts tend to generate significantly lower forecast errors. Such forecasting frameworks have the ability to outperform the ARIMA model, given the negative and significant coefficient. By contrast, the use of regression-based, futures-based, as well as, forecasts based on product spreads, does not seem to significantly outperform the forecasting accuracy of the ARIMA model.

The fact that the MIDAS models, as well as, the combined forecasts tend to produce

the lower forecast errors can be interpreted as follows. Primarily, it suggests that oil price forecasts are still heavily impacted by the fundamental factors of the oil market (i.e., unanticipated changes in oil supply, oil demand, and inventory levels). Obviously, this is something that has been highlighted by the vast majority of the oil price forecasting papers. More importantly, though, our finding suggests that the use of higher-frequency financial data tend to improve oil price forecasts, as also shown by Degiannakis and Filis (2018), which is justified by the increased financialisation of the oil market in the recent years.

Turning our attention to the choice of the type of oil price, our results suggest that the forecasting exercises that use either the US refiners' acquisition cost (RAC) or the WTI crude oil price tend to report higher RRMSEs compared to those studies that use the Brent crude oil price. Alquist et al. (2013) argue that although the RAC can be used to approximate global oil price movements, it cannot be viewed as the indicative proxy for the price that US refineries paid for crude oil. In this regard, Baumeister et al. (2014) are supportive in favour of the WTI spot price. According to them, WTI is not subject to revisions and it is also available without delays, which is not the case when the RAC is considered. This could justify the higher forecast error of the RAC compared to the Brent prices.

Furthermore, the WTI crude is considered to be more volatile than the Brent, which makes it harder to be accurately predicted. Possible reasons can be found in the geographical area that they are produced and the transportation costs. Brent is extracted at sea and transferred by ships, which makes it to be less dependent on abrupt changes in transportation costs. By contrast, the WTI is drilled in landlocked regions and thus, its price is affected by both higher transportation costs as well as pipeline bottlenecks and higher storage constraints (Baumeister and Kilian, 2015). Therefore, the use of Brent seems to provide better forecasting performance. Such findings are also in accordance with Manescu and Van Robays (2014) and Degiannakis and Filis (2018) who propose the importance to use the Brent spot price. Even more, the WTI crude oil market has attracted the attention of the non-commercial investors, via its futures contracts. Indeed, the WTI has the most

liquid and actively traded futures contracts in the crude oil market compared with Brent (Buyuksahin et al., 2013). This explains why WTI is regarded as a valuable financial asset by energy traders. Such trading activity results in the higher volatility for the WTI, which could further explain the lower forecasting accuracy for this crude oil benchmark. Such finding has important implications for end-users of oil price forecasts. Let us assume that there is a Permian Basin operator who is interested in forecasting the price of oil to help guide current production decisions. It is apparent that the appropriate oil price measure to forecast is the price of WTI. In such case, the operator should be aware that her forecasting framework should be improved so as to accommodate the fact that WTI is harder to predict.

As far as the difference between real and nominal oil prices, our evidence suggests that forecasts of real price tend to be better than forecasts of nominal one. A plausible explanation for this finding can be traced at the effect of inflation. More specifically, the higher forecast errors of the nominal oil prices could be explained by the fact that they have an inflation component, which adds uncertainty in the future path of oil prices. Put it differently, nominal oil price forecasts make also implicit assumptions about the future inflation, hence they are harder to predict.

Interestingly enough, we do not find evidence that the real-time forecasts are superior. Real-time forecasts are based on datasets that take into consideration delays in reporting relevant information or potential revisions in data series (for instance, this is particular relevant for oil production information). According to Baumeister et al. (2014) and Alquist et al. (2013), forecasters who ignore such constraints in the data series, tend to produce better forecasts. Nevertheless, our findings do not lend support to this claim.

Furthermore, forecasting horizon does systematically affect the forecasts. The variable 'horizon' appears to all models with a positive sign. This means that longer forecasting horizons produce higher RRMSE, indicating a lower forecast performance. Obviously, this is a plausible finding given that at longer horizons we expect the autoregressive and moving-average components of oil prices to prevail relatively to the fundamentals of the oil market or

the financial information. By contrast, we do not find any significant influence on the quality of forecasts from the forecasting period moderator, suggesting that either the more recent or the earlier forecasts in our dataset do not seem to exhibit different levels of predictive accuracy.

4.2 Robustness tests

Having analysed the first round results, it is important to use an array of robustness tests so as to verify the stability of our findings. The first test is to replace the Bayesian setting (BMA) with a frequentist one (FMA). So, we keep the basic rational of model averaging techniques. In this respect, the main difference between frequentist and Bayesian averaging is the construction of weights. Instead of using posterior model probabilities, the new weights are replaced with information criteria. In our exercise, we follow the approach proposed by Magnus et al. (2010) and extended by Amini and Parmeter (2012) who select the weights by minimising the Mallows criterion (Hansen, 2007). The main benefit is that this version of FMA is based on the orthogonalisation of the covariate space that leads to the significant reduction of the models that need to be estimated. In our case, the model space is not a problem due to the moderate number of regressors, as explained in the previous Sections.

The second test is to apply a pure frequentist least squares exercise without using any weighting scheme. Table 4 shows the results for both FMA and OLS with clustered standard errors at study level. Applying both types of frequentist analysis leads to results that are quantitatively and qualitatively similar to the BMA.

Overall, the moderator variables that were found to be robust drivers of the observed heterogeneity of the forecast errors remain the same; MIDAS models and combined forecasts tend to have a better performance. The opposite is true when the forecasting exercises are based on WTI and RAC price indexes. The horizon does continue to play a role, with longer periods resulting in worse forecasts. Finally, when the focus is on real prices, then these forecasts are superior compared to those based on nominal prices.

Table 4 here

Finally, we apply a variant of the least absolute shrinkage and selection operator (LASSO). This method combines the concept of minimising the least squares along with a shrinkage process that removes the drivers that are not important. The minimisation process can be written as: $(\psi_{kl}^{m} - \sum_{S=1}^{10} \gamma_S Z_{S,kl})^2 + \lambda \sum_{S=1}^{10} |\gamma|$, where λ is the shrinkage parameter. Even though the number of regressors is not very large, the insertion of a shrinkage parameter provides a natural way to test the robustness of model averaging. Here, we adopt its more widely used Bayesian version. Under certain assumptions regarding the prior distributions, the outcome is a set of estimations for those γ coefficients that have not being shrunk to zero.[6] Therefore, the variables that still have a non-zero coefficient after the shrinkage process are the most robust drives that explain the forecasting ability. The results are shown in Table 5. As in the frequentist exercise, MIDAS models and combined forecasts continue to report lower forecasts errors. In a similar vein, the use of the Brent price tends along with using real prices to provide better forecasts across the examined literature. Once more, better forecasts come from shorter forecasting horizons.

Table 5 here

5 Conclusions

The aim of this paper is to provide a comprehensive assessment of the factors that contribute to improve oil price forecasts, by conducting a meta-analysis. The timeliness of the paper stems from the fact that since the early 2000 and the regime change in oil price fluctuations, there is an ever increased interest in oil price forecasts. However, despite the numerous efforts, there is no empirical evidence of relevant studies to summarise the key factors that contribute to the accuracy of such forecasts. Thus, our quantitative survey on models contributes to the practice of oil price forecasting. To the best of our knowledge, this is the first study

[6] See Appendix 3 for technical details.

that evaluates the forecast accuracy performance of the forecasting models by focusing on the relative root mean squared error (RRMSE).

Using a dataset that covers a large range of different models, datasets, horizons as well as different oil price benchmarks, we identify the most importance drivers of forecasting accuracy. Based on the RRMSE, which is the most frequently used metric of forecasting performance, we summarise our findings as follows. Firstly, the choice of the model plays an important role; MIDAS modelling provides systematically better predictions than other model strategies. The same is true for empirical exercises that use combined forecasts. Secondly, the price benchmark is also an important factor. Our evidence indicates that the forecasting ability is improved when the Brent price is used. The opposite is true when the WTI or the RAC are used. Finally, shorter forecasting horizons, as well as, the use of real prices generate forecasts of greater accuracy. By contrast, the forecasting period and the real-time datasets are not important factors of the forecasting ability. Our findings remain unchanged under a set of different robustness tests.

The present paper can be a useful guide for future studies, as oil price fluctuations are important for numerous stakeholders (policy makers, the oil industry, multinational corporations, households and investors, among others) and thus, the efforts for their accurate forecasts are expected to intensify in the future. For the purposes of the current meta-analysis, we make use of one specific metric (RRMSE). Hence, an extension of the current study would be the consideration of alternative forecasting accuracy metrics, such as the directional accuracy. We leave these issues for future research.

References

Alquist, R. and Kilian, L. (2010). What do we learn from the price of crude oil futures? *Journal of Applied econometrics*, 25(4):539–573.

Alquist, R., Kilian, L., and Vigfusson, R. J. (2013). Forecasting the price of oil. In *Handbook*

of economic forecasting, volume 2, pages 427–507. Elsevier.

Amini, S. M. and Parmeter, C. F. (2012). Comparison of model averaging techniques: Assessing growth determinants. *Journal of Applied Econometrics*, 27(5):870–876.

Baumeister, C. (2014). The art and science of forecasting the real price of oil. *Bank of Canada Review*, 2014(Spring):21–31.

Baumeister, C., Guérin, P., and Kilian, L. (2015). Do high-frequency financial data help forecast oil prices? the midas touch at work. *International Journal of Forecasting*, 31(2):238–252.

Baumeister, C. and Kilian, L. (2012). Real-time forecasts of the real price of oil. *Journal of Business & Economic Statistics*, 30(2):326–336.

Baumeister, C. and Kilian, L. (2014). What central bankers need to know about forecasting oil prices. *International Economic Review*, 55(3):869–889.

Baumeister, C. and Kilian, L. (2015). Forecasting the real price of oil in a changing world: a forecast combination approach. *Journal of Business & Economic Statistics*, 33(3):338–351.

Baumeister, C., Kilian, L., and Lee, T. K. (2014). Are there gains from pooling real-time oil price forecasts? *Energy Economics*, 46:S33–S43.

Baumeister, C., Kilian, L., and Zhou, X. (2018). Are product spreads useful for forecasting oil prices? an empirical evaluation of the verleger hypothesis. *Macroeconomic Dynamics*, 22(3):562–580.

Benmoussa, A. A., Ellwanger, R., and Snudden, S. (2020). The new benchmark for forecasts of the real price of crude oil. Technical report, Bank of Canada.

Bernanke, B. S. (2005). Remarks by governor ben s. bernanke at the sandridge lecture, virginia association of economics, richmond, virginia.

Buyuksahin, B., Lee, T. K., Moser, J. T., and Robe, M. A. (2013). Physical markets, paper markets and the wti-brent spread. *The Energy Journal*, 34(3).

Chinn, M. D. and Meese, R. A. (1995). Banking on currency forecasts: how predictable is change in money? *Journal of International Economics*, 38(1-2):161–178.

Coppola, A. (2008). Forecasting oil price movements: Exploiting the information in the futures market. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, 28(1):34–56.

Degiannakis, S. and Filis, G. (2018). Forecasting oil prices: High-frequency financial data are indeed useful. *Energy Economics*, 76:388–402.

Dimitropoulos, A., Oueslati, W., and Sintek, C. (2018). The rebound effect in road transport: A meta-analysis of empirical studies. *Energy Economics*, 75:163–179.

ECB (2015). Forecasting the oil of price. *ECB Economic Bulletin*, 4:87–98.

Eickmeier, S. and Ziegler, C. (2008). How successful are dynamic factor models at forecasting output and inflation? a meta-analytic approach. *Journal of Forecasting*, 27(3):237–265.

Fratzscher, M., Schneider, D., and Van Robays, I. (2014). Oil prices, exchange rates and asset prices.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4):1175–1189.

Havranek, T. and Kokes, O. (2015). Income elasticity of gasoline demand: A meta-analysis. *Energy Economics*, 47:77–86.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kilian, L. (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review*, 99(3):1053–69.

Kilian, L. (2010). Oil price volatility: Origins and effects. Technical report, WTO Staff Working Paper.

Kilian, L. and Murphy, D. P. (2014). The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics*, 29(3):454–478.

Knetsch, T. A. (2007). Forecasting the price of crude oil via convenience yield predictions. *Journal of Forecasting*, 26(7):527–549.

Koop, G. (2017). Bayesian methods for empirical macroeconomics with big data. *Review of Economic Analysis*, 9(1):33–56.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Magnus, J. R., Powell, O., and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics*, 154(2):139–153.

Manescu, C. and Van Robays, I. (2014). Forecasting the brent oil price addressing time-variation in forecast performance. *ECB Working Paper 1735*.

Mattmann, M., Logar, I., and Brouwer, R. (2016). Hydropower externalities: A meta-analysis. *Energy Economics*, 57:66–77.

Murat, A. and Tokat, E. (2009). Forecasting oil price movements with crack spread futures. *Energy Economics*, 31(1):85–90.

Naser, H. (2016). Estimating and forecasting the real prices of crude oil: A data rich model using a dynamic model averaging (dma) approach. *Energy Economics*, 56:75–87.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques*.

Table 1: Forecast errors across studies

Study	Mean	SD	95% CI	
1. Alquist and Kilian (2010)	1.034	0.009	1.015	1.053
2. Alquist et al. (2013)	0.952	0.007	0.938	0.966
3. Baumeister et al. (2015)	1.044	0.008	1.027	1.061
4. Baumeister and Kilian (2012)	0.877	0.004	0.867	0.886
5. Baumeister and Kilian (2014)	0.957	0.004	0.949	0.966
6. Baumeister and Kilian (2015)	0.942	0.017	0.907	0.977
7. Baumeister et al. (2014)	1.003	0.003	0.995	1.010
8. Baumeister et al. (2018)	0.994	0.004	0.985	1.002
9. Chen (2014)	1.109	0.015	1.078	1.139
10. Coppola (2008)	0.949	0.005	0.939	0.960
11. Degiannakis and Filis (2018)	0.995	0.006	0.983	1.007
12. Funk (2018)	0.947	0.002	0.943	0.951
13. Garratt et al. (2019)	0.975	0.050	0.875	1.074
14. Knetsch (2007)	0.852	0.010	0.830	0.873
15. Naser (2016)	0.977	0.003	0.971	0.984
16. Pak (2018)	1.140	0.017	1.105	1.175
17. Rubaszek (2021)	0.895	0.008	0.879	0.912
18. Snudden (2018)	1.251	0.042	1.167	1.335
19. Wang et al. (2015)	1.130	0.031	1.068	1.191
20. Wang et al. (2017)	0.957	0.002	0.952	0.962
21. Zhang et al. (2019)	1.137	0.051	1.036	1.237

Notes: The table reports the mean, the standard deviation (SD) as well as the 5th and 95th percentile values of the relative root mean squared errors (ψ) for different subsets of data.

Table 2: List of Moderator Variables

Variable Name	Description	Mean	SD
Forecasting Model			
Structural	1 if ψ_{kl} from a structural model	1.036	0.009
MIDAS	1 if ψ_{kl} from a MIDAS model	0.926	0.007
Regression	1 if ψ_{kl} from a regression-based model	1.011	0.005
Combined	1 if ψ_{kl} from a combined forecast model	0.924	0.003
Future	1 if ψ_{kl} from a futures-based model	1.053	0.009
Product	1 if ψ_{kl} from a product spread-based model	1.022	0.006
Oil Price			
WTI	1 if ψ_{kl} uses West Texas Intermediate	1.009	0.005
RAC	1 if ψ_{kl} uses refiners' acquisition cost	1.014	0.005
Real	1 if ψ_{kl} uses real price	0.997	0.003
Forecasting Features			
Horizon	Number of months	9.924	0.124
Period	Standardized date of the end-of-sample	8.406	0.071
Real-time	1 if ψ_{kl} uses real-time forecasts	1.003	0.003

Notes: The table shows the descriptive statistics of the variables that are detected as potential drivers of the RRMSEs' heterogeneity.

Table 3: Bayesian Model Averaging results

	BMA_1			BMA_2		
Variable	PIP	post Mean	post SD	PIP	post Mean	post SD
Forecasting Model						
Structural	0.048	0.001	0.004	0.104	0.002	0.007
MIDAS	0.981^b	-0.057	0.016	0.977^b	-0.057	0.016
Regression	0.029	-0.001	0.003	0.055	-0.001	0.004
Combined	0.999^a	-0.083	0.009	0.999^a	-0.082	0.010
Future	0.086	0.003	0.013	0.123	0.005	0.014
Product	0.039	0.001	0.004	0.091	0.002	0.006
Oil Price						
WTI	0.999^a	0.049	0.009	0.999^a	0.049	0.009
RAC	0.999^a	0.058	0.010	0.999^a	0.058	0.010
Real	0.951^b	-0.072	0.023	0.954^b	-0.072	0.023
Forecasting Features						
Horizon	0.999^a	0.003	0.000	0.999^a	0.003	0.000
Period	0.001	0.006	0.001	0.001	0.006	0.001
Real-time	0.084	0.001	0.004	0.143	0.002	0.006

Notes: PIP stands for posterior inclusion probability. For BMA_1 unit information prior is used as parameters' prior and uniform model prior is used as model prior. For BMA_2 hyper-g prior and beta-binomial are used as paremeter and model priors, respectively. a/b/c denotes decisive/strong/positive evidence of a regressor having an effect, respectively, according to Kass and Raftery (1995).

Table 4: Frequentist Model Averaging and Least Squares

	FMA		OLS	
Variable	Coefficient	SD	Coefficient	SD
Forecasting Model				
Structural	-0.039	0.041	0.027	0.056
MIDAS	-0.039*	0.032	-0.042	0.045
Regression	-0.019	0.029	0.002	0.046
Combined	-0.033*	0.024	-0.067**	0.033
Future	0.000	0.021	0.033	0.042
Product	0.001	0.018	0.026	0.041
Oil Price				
WTI	0.035*	0.022	0.049*	0.018
RAC	0.047*	0.037	0.057**	0.027
Real	-0.075*	0.044	-0.065	0.044
Forecasting Features				
Horizon	0.004*	0.000	0.003***	0.001
Period	0.000	0.001	0.006	0.004
Real-time	0.070	0.114	0.011	0.018

Notes: For the OLS estimates, clustered standard errors at study level are reported. ***, ** and * indicate statistical significant at 1%, 5% and 10%, respectively. For the case of FMA the asterisk is used for illustrative purposes only and should be cautiously interpreted as the results from this method do not correspond to only one specification, but they represent an average.

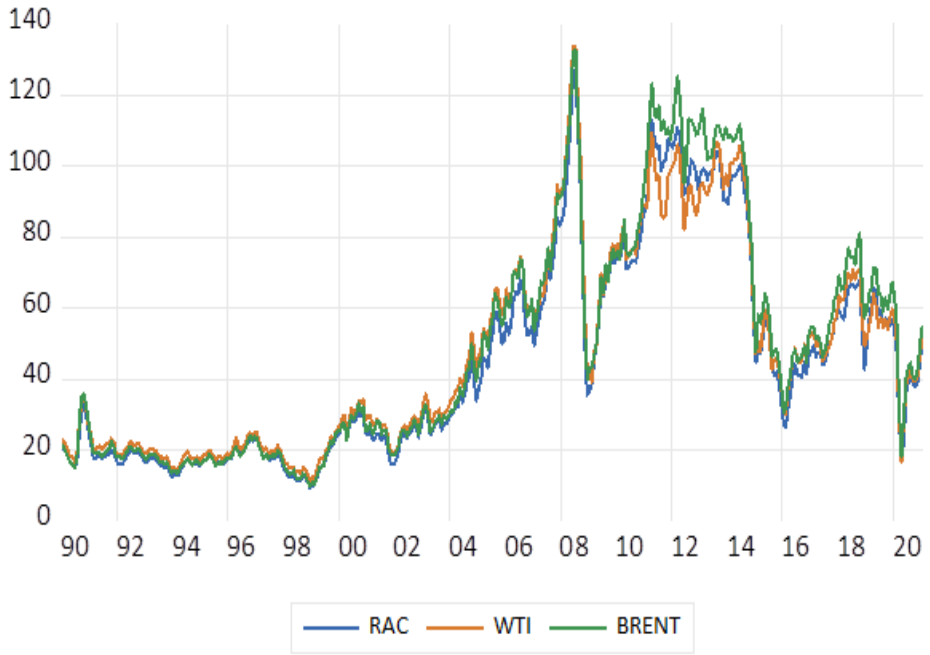
Table 5: LASSO estimates

Variable	post Mean	post SD	post τ
Forecasting Model			
Structural	0.000	0.000	0.000
MIDAS	-0.032✓	0.002	0.304
Regression	0.000	0.000	0.000
Combined	-0.092✓	0.004	0.401
Future	0.000	0.001	0.004
Product	0.000	0.000	0.000
Oil Price			
WTI	0.034✓	0.004	0.185
RAC	0.022✓	0.003	0.155
Real	0.029✓	0.004	0.158
Forecasting Features			
Horizon	0.015✓	0.004	0.011
Period	0.000	0.000	0.000
Real-time	0.001	0.001	0.001

Notes: ✓indicates the variables whose coefficients remain non-zero after the shrinkage process. Posterior τ is referring to mean of the posterior distribution of the hyperparameter τ that determines the variance of γ parameters. Details are explained in the appendix 3.

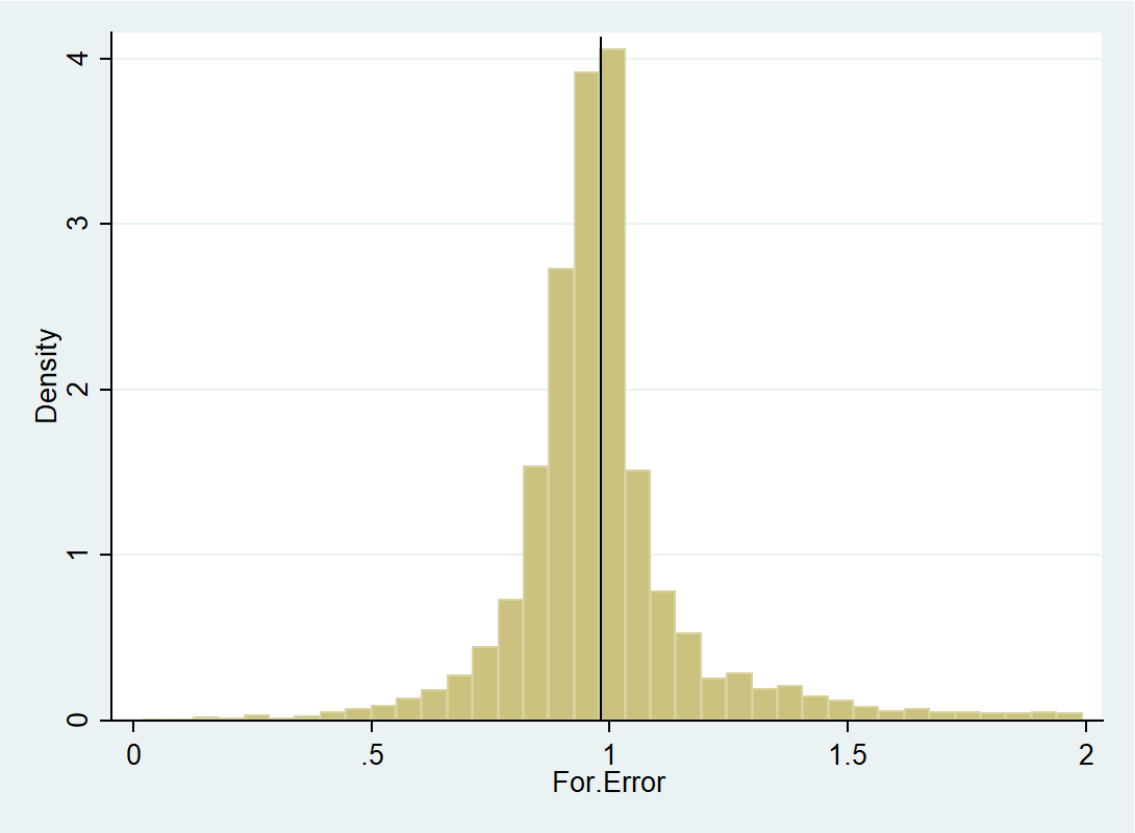
Figures

Figure 1: Nominal (spot) oil price



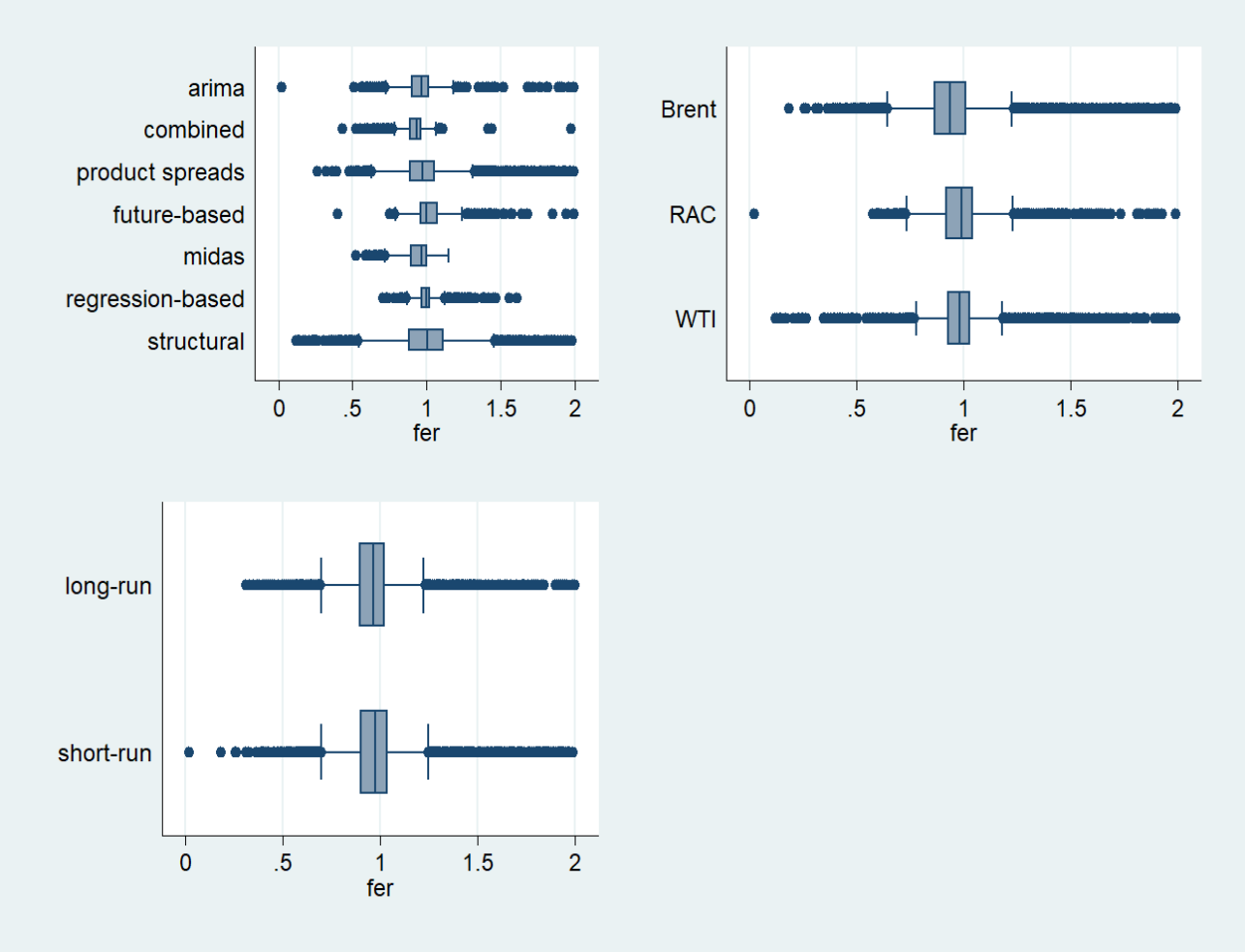
Notes: The figure demonstrates monthly data for the nominal US refiners' acquisition cost of crude oil imports (RAC), the nominal West Texas Intermediate (WTI) price and the nominal Brent (BRENT) price. The time period spans from January 1990 to January 2021. Source of collecting the data: Energy Information Administration.

Figure 2: Histogram of RRMSEs



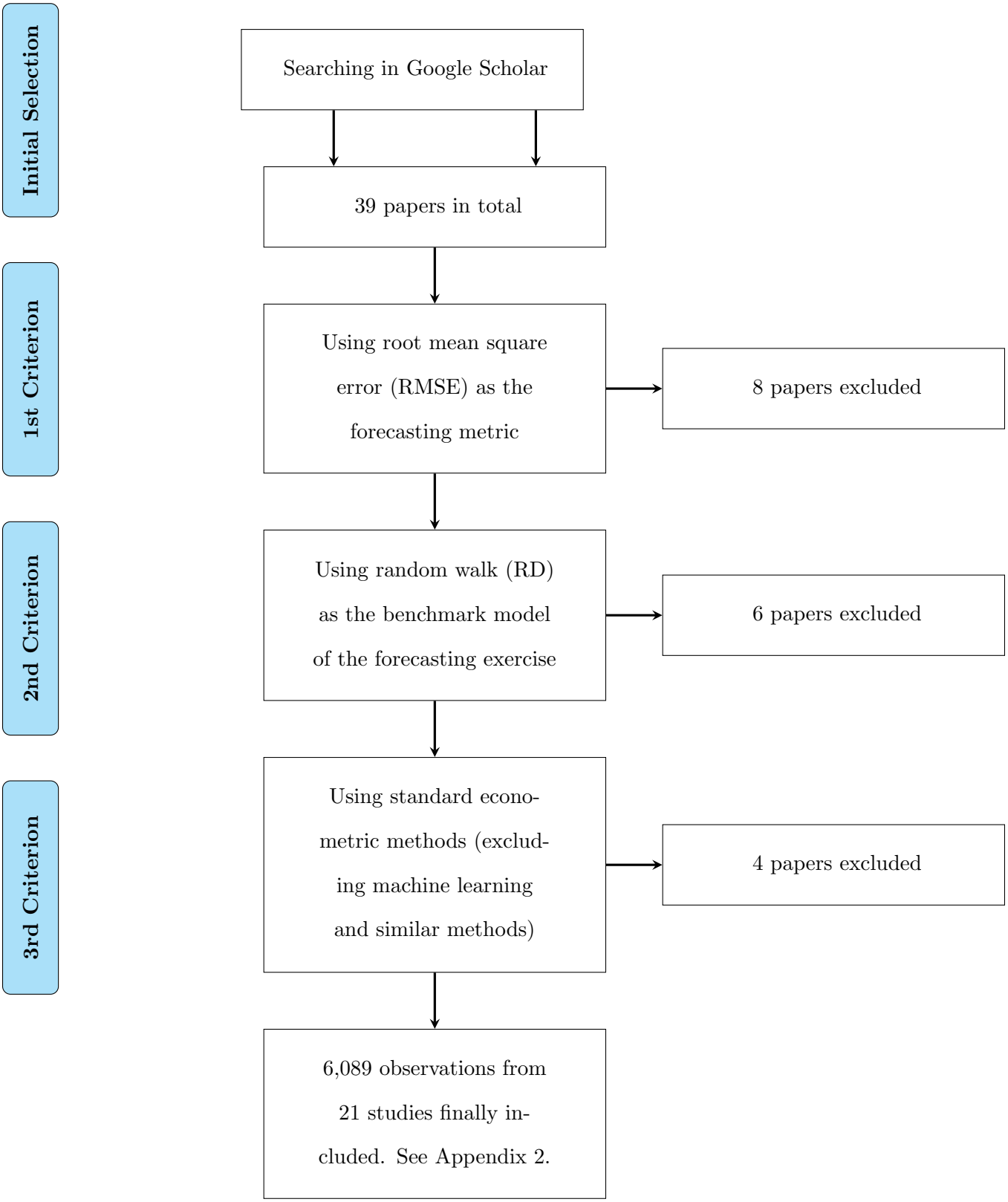
Notes: The figure depicts the histogram of the relative forecasting root mean squared errors, ψ reported by individual studies. The black solid line indicates the sample mean.

Figure 3: Heterogeneity of RRMSEs across different forecasting models, prices and horizons



Notes: The upper left boxplot shows the heterogeneity of the reported relative root mean squared errors (RRMSEs) across different models. The upper right boxplot shows the heterogeneity of RRMSEs across different oil prices. The bottom boxplot shows the heterogeneity of the RRMSEs across different forecasting horizons.

Appendix 1: PRISMA chart



Appendix 2: Studies used

Alquist, R. and Kilian, L. (2010). What do we learn from the price of crude oil futures? *Journal of Applied Econometrics*, 25(4):539-573.

Alquist, R., Kilian, L., and Vigfusson, R. J. (2013). *Forecasting the price of oil.* In Handbook of Economic Forecasting, volume 2, pages 427-507. Elsevier.

Baumeister, C., Guerin, P., and Kilian, L. (2015). Do high-frequency financial data help forecast oil prices? the MIDAS touch at work. *International Journal of Forecasting*, 31(2):238-252.

Baumeister, C. and Kilian, L. (2012). Real-time forecasts of the real price of oil. *Journal of Business Economic Statistics*, 30(2):326-336.

Baumeister, C. and Kilian, L. (2014). What central bankers need to know about forecasting oil prices. *International Economic Review*, 55(3):869-889.

Baumeister, C. and Kilian, L. (2015). Forecasting the real price of oil in a changing world: a forecast combination approach. *Journal of Business Economic Statistics*, 33(3):338-351.

Baumeister, C., Kilian, L., and Lee, T. K. (2014). Are there gains from pooling real-time oil price forecasts? *Energy Economics*, 46:S33-S43.

Baumeister, C., Kilian, L., and Zhou, X. (2018). Are product spreads useful for forecasting oil prices? an empirical evaluation of the Verleger hypothesis. *Macroeconomic Dynamics*, 22(3):562-580.

Chen, S.-S. (2014). Forecasting crude oil price movements with oil-sensitive stocks. *Economic Inquiry*, 52(2):830-844.

Coppola, A. (2008). Forecasting oil price movements: Exploiting the information in the futures market. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, 28(1):34-56.

Degiannakis, S. and Filis, G. (2018). Forecasting oil prices: High-frequency financial data are indeed useful. *Energy Economics*, 76:388-402.

Funk, C. (2018). Forecasting the real price of oil-time-variation and forecast combination. *Energy Economics*, 76:288-302.

Garratt, A. Vahey, S., Zhang, Y. (2019). Real-time forecast combinations for the oil price, *Journal of Applied Econometrics*, 34, 456-462.

Knetsch, T. A. (2007). Forecasting the price of crude oil via convenience yield predictions. *Journal of Forecasting*, 26(7):527-549.

Naser, H. (2016). Estimating and forecasting the real prices of crude oil: A data rich model using a dynamic model averaging (DMA) approach. *Energy Economics*, 56:75-87.

Pak, A. (2018). Predicting crude oil prices: Replication of the empirical results in "What do we learn from the price of crude oil?. *Journal of Applied Econometrics*,33,160-163.

Rubaszek, M. (2021). Forecasting crude oil prices with DSGE models, *International Journal of Forecasting*, 37, 531–546.

Snudden, S. (2018). Targeted growth rates for long-horizon crude oil price forecasts,*International Journal of Forecasting*, 34, 1-16.

Wang, Y., Liu, L., Diao, X., and Wu, C. (2015). Forecasting the real prices of crude oil under economic and statistical constraints. *Energy Economics*, 51:599-608.

Wang, Y., Liu, L., and Wu, C. (2017). Forecasting the real prices of crude oil using forecast combinations over time-varying parameter models. *Energy Economics*, 66:337-348.

Zhang, Y.-J. and Wang, J.-L. (2019). Do high-frequency stock market data help forecast crude oil prices? evidence from the MIDAS models. *Energy Economics*, 78:192-201.

Appendix 3: Technical details

The key concept is that BMA treats each model, M_i as extra parameters that their posterior have to be estimated. Starting from Bayes rule:

$$p(M_i|\psi, Z) \propto p(\psi|M_i, Z)p(M_i) \tag{A.1}$$

The left-hand side term is the posterior model probability (PMP), while the right-hand side term is the marginal likelihood function times the prior probability of model M_i.

$$p(\gamma|\psi, Z) = \sum_{j=1}^{1024} p(\gamma_j|\psi, Z, M_j)p(M_j|\psi, Z) \tag{A.2}$$

where $p(\gamma_j|\psi, Z, M_j)$ is the posterior distribution under model M_j and $p(M_j|\psi, Z)$ is the posterior model probability. The above equation shows that the posterior model probabilities are used as weights. More precisely, the posterior density of γ for each model M_i is weighted by the posterior model probability of each model M_i.

As far as the parameters priors are concerned, we follow Zellner (1986) and assume that the variance is proportional to

$$\gamma_i|g \sim N(0, \sigma^2(g(X_iX_i)^{-1}) \tag{A.3}$$

In this study, we employ two different choices regarding g. Firstly, we set $g = N$, which leads to the most trivial case of unit information prior (UIP), where N is the sample size. Secondly, we employ the hyper-g prior as suggested by Liang et al. (2008). Specifically, $\frac{g}{1+g}$ $\sim Beta(1, \frac{a}{2} - 1)$, where $\alpha \in$ (2,4] with a Beta distribution mean equal to $\frac{2}{\alpha}$. Regarding the model priors, we assume the binomial model prior according to which the model probability is given by $p(M_j) = \delta^{k_j}(1-\delta)^{K-k_j}$, where K is the maximum number of regressors, k_j is the number of regressors included in the model M_j and δ is a hyperparameter that expresses the probability of each regressor. Based on this assumption, we discern between two different

cases. Firstly, we set $\delta=\frac{1}{2}$ assigning equal probability to all models under consideration. Secondly, we use an alternative model prior that is less restrictive as far as the model size is concerned, assuming a hyperprior beta-binomial.

The LASSO was originally developed as a frequentist method for variable selection. Assuming the following distributions for the parameters of interest:

$$\gamma_i|h, \tau_i^2 \sim N(0, h^{-1}\tau_i^2) \tag{A.4}$$

$$p(\tau_i^2|\lambda) \propto exp(-\lambda\tau_i^2/2) \tag{A.5}$$

$$p(h) \propto 1/h \tag{A.6}$$

where h and τ are the crucial hyperparameters that determine the variance of the parameters of interest (γ), the inference is based on simulations on four posterior conditional distributions: $p(\gamma|\psi, h, \tau^2, \lambda^2)$, $p(h|\psi, \gamma, \tau^2, \lambda^2)$, $p(\tau^2|\psi, \gamma, h, \lambda^2)$ and $p(\lambda^2|\psi, \gamma, h, \tau^2)$. Following the standard practice, we assume an independent normal-gamma prior for the first two conditionals. For h we assume a non-informative prior as indicated by equation (A.6), while for τ (and for calculative purposes for $1/\tau$ we assume an inverse gamma and for λ we assume a gamma distribution. With this prior structure, the simulations are reduced to a Gibbs sampler. In the main part of the paper, we report the posterior mean and the standard deviation for γ and the posterior mean for τ that shows which variables can remain in the model and which ones can be deleted. Coefficients that their estimated τ is zero are deleted from the model as their estimated mean is zero as equation (A.4) indicates. On the other hand, the meta-regression model keeps only the determinants whose estimated γ_i have a variance that is larger than zero (i.e., $\tau_i > 0$).