ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Econometrics & Management Science: Quantitative Finance

Forecasting Non-Causal Processes

Master's Thesis

September 9, 2018

Author
E. Streithorst
369981

Supervisor
A. Pick

Co-Reader
A.A. Naghi

Abstract

This research explores the forecast quality of a non-causal autoregressive model compared to existing causal models. The forecasted data consists of electricity, gas and oil prices. Different degrees of freedom in the error term of the non-causal model and different forecast horizons, up to 12 months, are tested. Overall the non-causal forecasts perform comparable to causal and vector autoregressive model forecasts according to Diebold Mariano tests for electricity and gas data. For oil data the non-causal forecasts perform worse. Lower degrees of freedom are favored in the non-causal models. Parameter estimations converge in the non-causal model, therefore the results are robust.

Keywords: Non-causal, Forecasting, Economic Time Series, Autoregressive, Vector Autoregressive, Sampling Importance Resampling, Diebold-Mariano test

I would like to thank Andreas Pick. Although our meetings were far and few between, they were nonetheless very helpful. I would also like to thank my family and their eagerness to follow my progress, and my friends with whom I could always share my thoughts and concerns when I needed to.

Never late, nor early, but precisely when it was meant to be.

Gorinchem, July 2018

Erik Streithorst

Contents

List of Tables

List of Figures

1 Introduction

Forecasting economic time series is a popular topic in financial academic literature. The voluminous research is a direct proof of this. There are many different ideas of how an economic time series behaves and what its drivers are, hence many different approaches have been examined such as the use of autoregressive (AR) models (Campbell et al., 1997) or the generalized autoregressive conditional heteroskedasticity (GARCH) models developed by Engle (1982) and 19 years later emphasized its importance in Engle (2001). More recent machine learning techniques are applied to model economic time series as well (J. V. Hansen et al., 2006).

Many different approaches have been used to model economic time series, however Lanne & Saikkonen (2011) state "all economic applications so far restrict themselves to causal autoregressive models where the current value of the variable of interest is forced to depend only on the present and past values of the errors of the model." (p.1). Following this statement it is a reasonable next step to then explore the use of non-causal models which include leads in addition to lags. Economic theory to support a non-causal approach is proposed by L. P. Hansen & Sargent (1991). The information set of an economic agent who operates in the sector of interest can be greater than that of the econometrician sitting behind a desk. This discrepancy gives rise to a nonfundamental solution to the modeling problem in the form of non-causality. Theoretically, this extra information allows the agents to forecast future values of an economic variable in question which is not possible for the econometrician. This results in a non-causal representation with predictable errors.

Non-causal models incorporate predictable errors by making assumptions on their distribution. Theoretically, forecasting performance can therefore be improved if the time series shows signs of non-causality. Problems do arise in identifying whether a time series is non-causal or not. If a time series shows signs of purely forward looking behavior, then a non-causal model is justified. The identification problem is solved by having a non-normal distribution of the error term when fitting an autoregressive model on the data (Weiss, 1975).

Theoretical research on non-causal processes is not unfamiliar but the topic is less popular than research on causal processes. For example Davis & Resnick (1986) and Rosenblatt (2000) provide a general approach to non-causal processes which is necessary for building the theoretical framework on which all following research relies. Other research on non-causal processes with a theoretical statistical approach includes, (Rosenblatt, 1995), (Gassiat, 1993) and (Huang & Pawitan, 2000)). Different research areas such as physics also makes use of these processes (Falnes, 1995). A reason for the popularity of causal processes is probably the more straightforward implementation as opposed to the non-causal processes, as there are less issues with time and unknown data.

This research compares forecasting performance of different (non-)causal models. The research question is: does a non-causal process improve forecasting accuracy of economic time series compared to causal processes? The non-causal models used in this research are non-causal autoregressive models. By using these specific type of models I stay close to the popular existing autoregressive methods, but incorporate non-causality. These models are denoted by AR(r,s) where r is the number of lags and s is the number of leads included. Recent literature of this topic has explored non-causal modeling (Lanne & Saikkonen, 2011) and forecasting (Gourieroux & Jasiak, 2016). The additional value of this research is the evaluation of forecast quality in an economic setting.

The economic time series consist of oil, gas and electricity prices measured on a monthly basis. The choice for the energy market has been made because I have not found previous literature covering the energy market with a non-causal approach and therefore this is an addition to existing research. A common way to model oil price is to use an AR model, for example Kilian (2008) captures exogenous shocks in its price. Baumeister & Peersman (2013) uses price elasticities to account for volatility changes. The usage of AR models benefits this research since they are compared to non-causal AR models. The setup of these two models is similar in their basis and therefore it is not unreasonable to compare their performance.

2 Data

This section describes the data that is used in this research. Section 1 mentions the use of non-causal processes in forecasting economic time series which are often found to be 'difficult' to forecast due to the presence of high variance and/or kurtosis. This research examines three such series: electricity, gas and oil prices. Electricity and gas data consists of monthly European harmonized consumer price index[1] from December 2009 to July 2017. Simple monthly growth rates[2] are taken of electricity and gas prices. The oil data is constructed by using the U.S. crude oil composite acquisition cost by refiners[3] and the U.S. CPI[4] monthly from January 1974 to July 2017. First index the acquisition cost with base 100 on January 1974. Next divide the price index by the corresponding CPI to get the real price of oil. The oil price is transformed by taking simple yearly growth rates[2]. By taking growth rates the first month of data is 'lost'. therefore growth rates for electricity and gas are available from January 2010 to July 2017 and for oil from February 1974 to July 2017. For all three time series the null hypothesis of a unit root is rejected according to an Augmented Dickey-Fuller test (Dickey & Fuller, 1979) and trend-stationarity is not rejected according to a KPSS test (Kwiatkowski et al.,

[1] data retrieved from *Centraal Bureau voor de Statistiek* (CBS)

[2] Simple monthly growth rate $\Delta y_t = \frac{y_t - y_{t-1}}{y_{t-1}} * 100$

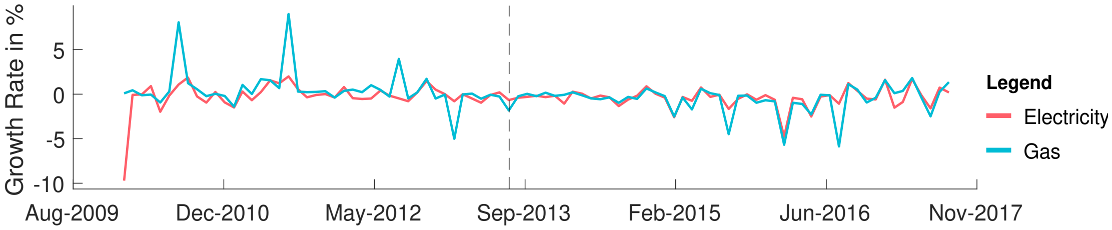
[3] data retrieved from U.S. Energy Information Administration

[4] data retrieved from U.S. Bureau of Labor Statistics

1992). Figure 1 shows the growth rates of electricity and gas prices, Figure 2 shows the growth rates of the real price of oil.

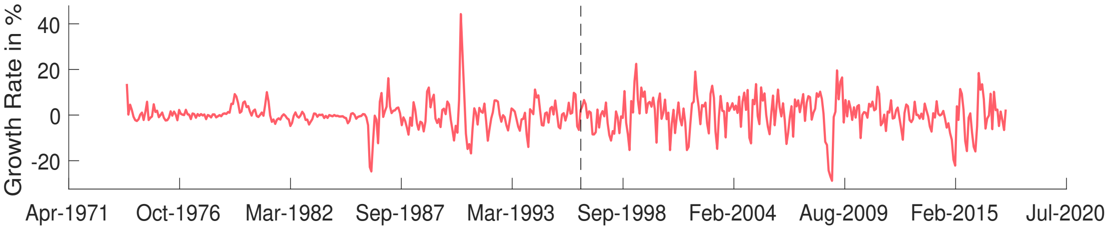
This research replicates part of the findings in Kilian (2009) and therefore uses some of its data as well. This data consists of the log differences of worldwide crude oil production[3], a real economic activity indicator custom made in Kilian (2009) and finally the real price of oil as described above without taking yearly growth rates and expressed in logs. A graphical overview of this data can be found in Figure 6 in appendix A. The three preceding time series are used in comparison with the real price of oil and is therefore available from January 1974 to July 2017 as well.

Figure 1: **Electricity and gas data**



Monthly growth rate of electricity and gas prices.

Figure 2: **Oil Data**



Monthly growth rate of real price of oil.

Table 1 contains summary statistics of the growth rates of electricity, gas and oil. Note that the skewness and kurtosis of electricity are relatively large compared to the other time series. The first growth rate from December 2009 to January 2010 of electricity is close to −10% which influences these outcomes heavily. Furthermore, the variance of oil is larger which is easily seen in figures 1 and 2. The electricity and gas data stays roughly between −5% and +5%, whereas oil is between −20% and +20%.

Table 1: **Summary Statistics**

	Electricity	Gas	Oil
Sample Period	01/2010 - 07/2017	01/2010 - 07/2017	02/1974 - 07/2017
Forecast Period	08/2013 - 07/2017	08/2013 - 07/2017	08/1996 - 07/2017
Mean	-0.33	-0.06	0.27
Variance	1.96	3.71	46.93
Skewness	-3.58	1.16	0.06
Kurtosis	24.48	12.01	8.21

Summary statistics of the three datasets used in this research. Sample Period denotes all available data, Forecast Period denotes the period that will be forecasted.

3 Models

This research uses causal and non-causal models. The quality of the non-causal models is of main interest and is measured by comparing them to causal models. I make a selection of causal models based on existing literature on electricity, gas and oil price forecasting. First, I introduce the non-causal AR(r,s) model. After this the AR(p), VAR and Local Level models follow.

3.1 Non-causal AR(r,s)

A first important issue before defining the non-causal AR(r,s) model, is to elaborate as to why this model is plausible to use in economic time series. If the time series shows signs of purely forward looking behavior, then a non-causal model is justified. This concept is also known in literature as a stochastic process which is not time-reversible. Usually in economic time series the Box-Jenkins approach is used to check the autocorrelation of a time series. Economic time series usually show signs of high autocorrelation and this has always been assumed to be evidence of causal behavior. The problem with this conclusion however is that a purely causal model with p lags corresponds to a purely non-causal model with p leads when this high autocorrelation is used as a persistence measure. This is because the underlying Gaussian assumption causes the spectral densities to be the same[5] for a non-causal model with coefficients $(\psi_1 L^{-1}, ..., \psi_n L^{-n})$ and a causal model with coefficients $(\psi_1 L, ..., \psi_n L^n)$ where L is the lag operator. Intuitively this correspondence can be explained by the symmetry of the Gaussian probability distribution. If one were to take independent draws from a Gaussian distribution over time, the draws will covary symmetrically at times $t + h$ and $t - h$ or in other words, it is time-reversible. Weiss (1975) proofs that if $X(t)$ is a stationary Gaussian process then $X(t)$ is time-reversible. A partial converse theorem is

[5] The spectral density of y_t in (1) is $\frac{\sigma^2}{2\pi}\left|\phi(e^{-i\omega})\psi(e^{-i\omega})\right|^2$

proven as well: if $X(t)$ is a stationary, time-reversible, ARMA process then the error term, say ε_t, is normally distributed. Therefore to use a non-causal process we should be able to identify it. This is possible when the process is not time-reversible. Following the theorems of Weiss (1975) a way of doing this is when we find a non-normal error term after fitting an AR(p) model to the data. To check if this is the case I use the partial autocorrelation function to fit an appropriate AR(p) model to each of the time series and test the resulting residuals for normality using the Jarque-Bera test (1980). All residuals were found to be non-normal and thus a non-causal model is justified.

The non-causal AR(r,s) model is used to fit on the following stochastic process y_t generated by

$$\Phi(L)\Psi(L^{-1})y_t = \varepsilon_t, \tag{1}$$

where $\Phi(L) = 1 - \phi_1 L - ... - \phi_r L^r$ and $\Psi(L^{-1}) = 1 - \psi_1 L^{-1} - ... - \psi_s L^{-s}$ with roots strictly outside the unit circle. The error term ε_t is i.i.d. such that $E(|\varepsilon_t|^{\delta}) < \infty$ for $\delta > 0$. r and s can take on any non-negative integer value where three cases are 1.) $r = 0$ and $s > 0$ results in a purely non-causal model with only leads and 2.) $r > 0$ and $s = 0$ results in a purely causal model with only lags and 3.) $r > 0$ and $s > 0$ results in a mixed model.

Following the preceding reasoning, throughout this entire research I assume the error term ε_t in the AR(r,s) model to be non-Gaussian and to have a Lebesgue density $f_s(x;\lambda) = \sigma^{-1} f(\sigma^{-1}x;\lambda)$ with parameter vector λ and scale parameter σ. Specifically I choose a t-distribution with ν degrees of freedom. This distribution is not unfamiliar in time series modeling because of its higher kurtosis (*fat tails*) than the Gaussian distribution. Kilian (2009) and Gourieroux & Jasiak (2016) use this distribution with $\nu = 3$ and $\nu = 1$ degrees of freedom respectively.

The structure of the AR(r,s) model when choosing $r = s = 1$ is as follows

$$\begin{aligned}
(1 - \phi_1 L)(1 - \psi_1 L^{-1})y_t &= \varepsilon_t \\
(1 + \phi_1\psi_1)y_t &= \phi_1 y_{t-1} + \psi_1 y_{t+1} + \varepsilon_t \\
y_t &= \frac{\phi_1}{1 + \phi_1\psi_1} y_{t-1} + \frac{\psi_1}{1 + \phi_1\psi_1} y_{t+1} + \frac{\varepsilon_t}{1 + \phi_1\psi_1}.
\end{aligned}$$

Writing the complete process for larger values of r and s quickly becomes unclear as it includes $r + s$ different terms of y and the coefficients contain cross-terms of ϕ and ψ.

In the data generating process of Equation (1) a causal and a non-causal component can be distinguished. These components are u_t and v_t respectively. The causal component

u_t and its moving average representation are defined as

$$\begin{aligned} u_t &= \Psi(L^{-1})y_t \\ \Phi(L)u_t &= \varepsilon_t \\ u_t &= \phi_1 u_{t-1} + ... + \phi_r u_{t-r} + \varepsilon_t \\ u_t &= \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j}, \end{aligned} \tag{2}$$

where $\alpha_0 = 1$ and the coefficients α_j converge to zero as $j \to \infty$. Note that the final step includes repetitive substitution of all lagged values of u. This will create multiple cross-terms in front of the summation sign which all contain powers of ϕ and therefore converge to zero due to the stationarity of the process. In a similar fashion derive the non-causal component v_t and its moving average representation

$$\begin{aligned} v_t &= \Phi(L)y_t \\ v_t &= \sum_{j=0}^{\infty} \beta_j \varepsilon_{t+j}, \end{aligned} \tag{3}$$

where $\beta_0 = 1$ and the coefficients β_j converge to zero as $j \to \infty$, again due to the cross-terms of ψ. Combining (2) and (3) to compute the moving average representation for y_t gives

$$y_t = \sum_{j=-\infty}^{\infty} \gamma_j \epsilon_{t-j}, \tag{4}$$

where $\gamma_0 = 1$ and the coefficient γ_j converges to zero as $|j| \to \infty$.

To be able to estimate the parameters of the model, determine the log-likelihood function $l_T(\theta)$ of y_t. To do this, first note that the following information sets are equivalent.

i. $(y_1, ..., y_T)$
ii. $(y_1, ..., y_r, v_{r+1}, ..., v_T)$
iii. $(u_1, ..., u_{T-s}, y_{T-s+1}, ..., y_T)$
iv. $(y_1, ..., y_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, v_{T-s+1}, ..., v_T)$
v. $(u_1, ..., u_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, y_{T-s+1}, ..., y_T)$
vi. $(u_1, ..., u_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, v_{T-s+1}, ..., v_T)$,

where ii. is obtained by using (3) on i. and similarly, iii. by using (2). Obtain iv. by combining i. and ii. and using (1), similarly obtain v. by combining i. and iii. Finally vi. is a combination of iv. and v. The equivalence of these information sets implicates there exists a transformation between each. Lanne & Saikkonen (2011) derive this transformation through the following matrix representations which can be constructed by using equations (2) and (3). In these representations the vectors $(u_1, ..., u_{T-s}, v_{T-s+1}, ..., v_T)$

and $(u_1, ..., u_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, \upsilon_{T-s+1}, ..., \upsilon_T)$ are referred to as x and z respectively.

$$\begin{bmatrix} u_1 \\ \vdots \\ u_{T-s} \\ \upsilon_{T-s+1} \\ \vdots \\ \upsilon_T \end{bmatrix} = \begin{bmatrix} y_1 - \psi_1 y_2 - \cdots - \psi_s y_{s+1} \\ \vdots \\ y_{T-s} - \psi_1 y_{T-s+1} - \cdots - \psi_s y_T \\ y_{T-s+1} - \phi_1 y_{T-s} - \cdots - \phi_r y_{T-s+1-r} \\ \vdots \\ y_T - \phi_1 y_{T-1} - \cdots - \phi_r y_{T-r} \end{bmatrix} = A \begin{bmatrix} y_1 \\ \vdots \\ y_{T-s} \\ y_{T-s+1} \\ \vdots \\ y_T \end{bmatrix} \tag{5}$$

$$\begin{bmatrix} u_1 \\ \vdots \\ u_r \\ \varepsilon_{r+1} \\ \vdots \\ \varepsilon_{T-s} \\ \upsilon_{T-s+1} \\ \vdots \\ \upsilon_T \end{bmatrix} = \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ u_{r+1} - \phi_1 u_r - \cdots - \phi_r u_1 \\ \vdots \\ u_{T-s} - \phi_1 u_{T-s-1} - \cdots - \phi_r u_{T-s-r} \\ \upsilon_{T-s+1} \\ \vdots \\ \upsilon_T \end{bmatrix} = C \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ u_{r+1} \\ \vdots \\ u_{T-s} \\ \upsilon_{T-s+1} \\ \vdots \\ \upsilon_T \end{bmatrix}, \tag{6}$$

where (5) and (6) are $x = Ay$ and $z = Cx$ respectively. These matrix representations give the following relationship between y and z: $z = CAy$.

The following derivations use the density function of z and finally result in the log-likelihood function of y_t (Lanne & Saikkonen, 2011). The density function of z is equal to the density function of y multiplied by $|det(C)||det(A)|$. C is an upper triangular matrix with ones on the diagonal therefore $det(C) = 1$. The moving average representations of u_t and υ_t in (2) and (3) respectively, show that the elements $(u_1, ..., u_r), (\varepsilon_{r+1}, ..., \varepsilon_{T-s}), (\upsilon_{T-s+1}, ..., \upsilon_T)$ of z are independent, therefore define the density function of z as

$$h_u(u_1, ..., u_r)\left(\prod_{t=r+1}^{T-s} f_\sigma(\varepsilon_t; \lambda)\right) h_\upsilon(\upsilon_{T-s+1}, ..., \upsilon_T),$$

where h_u and h_υ are the joint density functions of $(u_1, ..., u_r)$ and $(\upsilon_{T-s+1}, ..., \upsilon_T)$ respectively. The density function of y then equals

$$\begin{aligned} h_u(\psi(L^{-1})y_1, ..., \psi(L^{-1})y_r)&\left(\prod_{t=r+1}^{T-s} f_\sigma(\phi(L)\psi(L^{-1})y_t; \lambda)\right) \\ &h_\upsilon(\phi(L)y_{T-s+1}, ..., \phi(L)y_T)|det(A)|. \end{aligned} \tag{7}$$

h_u, h_υ and $det(A)$ are independent of sample size T and therefore the second term of Equation (7) approximates the density. The approximated log-likelihood function of y_t

then becomes

$$l_T(\theta) = \sum_{T=r+1}^{T-s} g_t(\theta) = \sum_{T=r+1}^{T-s} \Big[log(f(\psi(L^{-1})\phi(L)y_t;\lambda)) - log(\sigma)\Big].$$

If the chosen t-distribution is plugged in, $\theta = (\phi, \psi, \sigma, \nu)$ and g_t is the probability density function (pdf) of the t-distribution, the resulting function is

$$l_T(\theta) = \sum_{t=r+1}^{T-s} log\left[\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\Big(1 + \frac{\sigma^{-2}(\psi(L^{-1})\phi(L)y_t)^2}{\nu}\Big)^{-\frac{\nu+1}{2}}\right] - log(\sigma). \tag{8}$$

Estimate the parameters in θ by maximizing the log-likelihood function with a *Quasi-Newton* approach. According to Gourieroux & Jasiak (2016) a recursive Berndt, Hall, Hall and Hausman (1974) algorithm requires less evaluations of the log-likelihood function and is therefore computationally less demanding. However the quality of the estimates is roughly the same, therefore I use the well-known Quasi-Newton approach. Table 2 provides an overview of estimated parameter values of all three datasets on which the model is estimated. The values of ϕ and ψ are persistent throughout different degrees of freedom (d.o.f.) with the exception of the AR(1,1) model on the oil data. The value of σ however increases when the d.o.f. increase. Higher d.o.f. causes the t-distribution to approach normality, the higher variance counters this effect. Recall that a normal error distribution is rejected by a Jarque-Bera test (1980). The values of the gas data are overall quite small, followed by electricity and oil with the greatest values. A possible reason for this could be the higher variance in the oil dataset, which causes the non-causal parameter to be of more influence.

Table 2: **AR(r,s) Parameter Estimates**

		Electricity				**Gas**				**Oil**			
		AR(0,1)	AR(0,2)	AR(0,3)	AR(1,1)	AR(0,1)	AR(0,2)	AR(0,3)	AR(1,1)	AR(0,1)	AR(0,2)	AR(0,3)	AR(1,1)
d.o.f. 1	ϕ	-	-	-	0.03	-	-	-	0.04	-	-	-	-0.05
	ψ	0.14	0.14	0.11	0.13	0.06	0.06	0.05	0.06	0.48	0.52	0.52	0.51
		-	0	-0.01	-	-	0.07	0.07	-	-	-0.06	-0.06	-
		-	-	0.14	-	-	-	0.17	-	-	-	0.01	-
	σ	0.44	0.43	0.4	0.42	0.39	0.39	0.35	0.38	2.48	2.46	2.46	2.45
d.o.f. 2	ϕ	-	-	-	0.03	-	-	-	0.05	-	-	-	-0.01
	ψ	0.13	0.14	0.12	0.12	0.07	0.06	0.06	0.06	0.46	0.5	0.5	0.47
		-	-0.02	-0.03	-	-	0.07	0.06	-	-	-0.08	-0.07	-
		-	-	0.14	-	-	-	0.14	-	-	-	-0.01	-
	σ	0.58	0.58	0.55	0.56	0.55	0.55	0.51	0.55	3.34	3.33	3.34	3.32
d.o.f. 3	ϕ	-	-	-	0.03	-	-	-	0.05	-	-	-	0.03
	ψ	0.13	0.14	0.12	0.12	0.07	0.07	0.06	0.06	0.45	0.5	0.5	0.43
		-	-0.03	-0.04	-	-	0.06	0.06	-	-	-0.09	-0.08	-
		-	-	0.14	-	-	-	0.12	-	-	-	-0.03	-
	σ	0.66	0.66	0.63	0.64	0.68	0.68	0.64	0.68	3.83	3.81	3.82	3.82
d.o.f. 5	ϕ	-	-	-	0.03	-	-	-	0.05	-	-	-	0.12
	ψ	0.13	0.14	0.13	0.12	0.08	0.07	0.07	0.06	0.45	0.5	0.5	0.35
		-	-0.04	-0.06	-	-	0.05	0.05	-	-	-0.1	-0.08	-
		-	-	0.14	-	-	-	0.11	-	-	-	-0.04	-
	σ	0.76	0.76	0.73	0.72	0.87	0.88	0.82	0.88	4.38	4.36	4.36	4.37
d.o.f. 10	ϕ	-	-	-	0.02	-	-	-	0.05	-	-	-	0.21
	ψ	0.13	0.14	0.15	0.12	0.08	0.08	0.07	0.06	0.46	0.51	0.51	0.29
		-	-0.05	-0.07	-	-	0.05	0.05	-	-	-0.11	-0.09	-
		-	-	0.15	-	-	-	0.12	-	-	-	-0.04	-
	σ	0.88	0.89	0.86	0.82	1.18	1.19	1.13	1.19	4.97	4.94	4.94	4.95

Values for parameter estimates of different setups of the AR(r,s) model. d.o.f. is the degrees of freedom used in the t-distribution in the log-likelihood. The horizontal dashed lines separate the lags, ϕ, leads, ψ and scale parameter σ.

3.2 Causal AR(p)

In context of this research the AR(p) model is simply a special case of the $AR(r, s)$ model where $r \geq 1$ and $s = 0$. The AR(p) model is defined as

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t, \tag{9}$$

where c is a constant, ϕ_i is the autoregressive parameter and the error term ε_t is white noise. One choice has to be made regarding the setup of this model, namely what number of lags to include. Two methods of choosing a lag order are the use of a partial autocorrelation function and the use of information criteria. In this research I choose to use the latter method. To choose a lag order I use two different information criteria, namely Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). Table 3 shows these values for five different AR(p) models. A lower value is favored, therefore the smallest value of every row is underlined. In addition a lower order model is favored, therefore I choose an AR(1) model for electricity and gas and an AR(2) model for oil.

Table 3: **Information Criteria**

		AR(1)	AR(2)	AR(3)	AR(4)	AR(5)
Electricity	AIC	318.2	317.2	317.9	319.2	317.9
	BIC	320.8	322.3	325.4	329.2	330.4
Gas	AIC	377.4	379.0	378.3	380.2	382.2
	BIC	379.9	384.0	385.8	390.2	394.7
Oil	AIC	3350	3343	3342	3342	3343
	BIC	3354	3352	3354	3359	3364

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values of different AR(p) models. Smallest values of each row are underlined.

Table 4 shows estimation results of the AR(p) models resulting from the chosen lag orders. We see that the constant is close to zero for all datasets which implies no positive or negative growth rate on average. The first coefficient is positive for every dataset which implies a certain persistence in growth rate, e.g. a positive growth rate in the previous month causes a positive push in the growth rate of this month and vice versa. Furthermore, Table 4 shows the test results of a Jarque-Bera test (Jarque & Bera, 1980). The null hypothesis of a normally distributed error is rejected for every dataset and indicates the possibility of a non-causal model to be identified (Section 3.1). For completeness appendix B provides all AR(p) parameter estimates.

Table 4: **AR(p) Estimation and Jarque-Bera Test Results**

	Model	Constant	ϕ_1	ϕ_2	Statistic	P-value
Electricity	AR(1)	-0.23	0.35	-	769	0.00
Gas	AR(1)	-0.05	0.11	-	325	0.00
Oil	AR(2)	0.16	0.55	-0.13	413	0.00

Estimation results of fitting an AR(p) model on full datasets. Constant, ϕ_1 and ϕ_2 are parameters of the model. Statistic and P-value refer to the Jarque-Bera test results of the residuals.

3.3 Causal VAR

The VAR model expands on the AR model incorporating lag dependency between different variables. The model is used in this research because Kilian (2009) shows promising forecasting results regarding the real price of oil. The VAR model of this research is therefore similar. It contains three variables that have been introduced in section 2, namely the percentage change in global crude oil production $\Delta prod_t$, the index of real economic activity rea_t and the real price of oil rpo_t. The general setup of this model is

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \varepsilon_t, \tag{10}$$

where all elements included are defined as follows.

$$y_t = \begin{pmatrix} \Delta prod_t \\ rea_t \\ rpo_t \end{pmatrix}, \quad c = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix}, \quad A_i = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad i = 1, ..., p, \quad \varepsilon_t = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{pmatrix},$$

where y, c and ε are 3×1 vectors and A is a 3×3 matrix. All elements in ε_t are assumed to be white noise. Kilian (2009) presents the model in structural form which defines a nonsingular matrix B_0. This matrix has the following two properties, $B_0^{-1} B_i = A_i$ and $B_0^{-1} w_t = \varepsilon_t$. The latter property implies that B_0 captures the impact effects of each of the structural shocks on each of the model variables. Equation (11) shows the structural VAR model

$$B_0 y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + w_t. \tag{11}$$

The number of lags included in the VAR model in this research is $p = 2$. The focus of this research is not on interpretation of structural shock effects, therefore the reduced-form VAR model in Equation (10) is used.

3.4 Local Level

The Local Level model (LLM) or otherwise known as a Random Walk with noise is one of the most basic approaches to modeling time series. Despite its simplicity it is a common benchmark for measuring performance of other models. Equations (12) and (13) show the Local Level model

$$y_t = \mu_t + \varepsilon_t \tag{12}$$

$$\mu_{t+1} = \mu_t + \eta_t, \tag{13}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and $\eta_t \sim \mathcal{N}(0, \sigma^2)$ for every value of t. The LLM can be interpreted as a state-space representation with (12) the observation equation and (13) the state equation. I will use this to derive the Kalman filter (1960) that will aid in forecasting in section 4.5.

4 Forecasting

This section describes how the models from section 3 are used to forecast prices. First every model will be estimated using a subsample of the sample period defined in Table 1, namely the period before the forecast period, then H values are forecasted. This process repeats itself throughout the entire forecast period, expanding the estimation window, until 07/2017 is reached. The setup of this section is as follows, first the choices that have been made in the overall forecasting process are described. Next, the succeeding subsections describe the difference in causal and non-causal forecasting and the forecasting procedures of the AR(r,s), AR(p), VAR and Local Level models.

Different forecasting horizons are used to compare model performance, namely 1, 2, 3, 6 and 12 months. This covers both short and medium term forecasting. Longer horizons of multiple years are not of interest to this research. This setup causes the first forecasted values to be made with a model that is estimated with less data than the final forecasted values when nearing the end of the sample period. Especially with the data on electricity and gas prices, the first estimation only uses 43 months of data. I expect the results of all models improve over time as the sample period increases.

4.1 Causal Versus Non-causal Forecasting

Forecasting with causal models is for most of the common models roughly the same, that is, the value y_t is only affected by previous values of y, innovation ε_t and possibly by some lagged innovations as well. A good example of this is a causal moving average model. This implies that the conditional expectation $E_t(y_{t+1})$ has only one unobserved innovation, ε_{t+1}. All other innovations are observed, albeit possibly not identified. A non-causal model however is affected by future innovations. To illustrate this, use

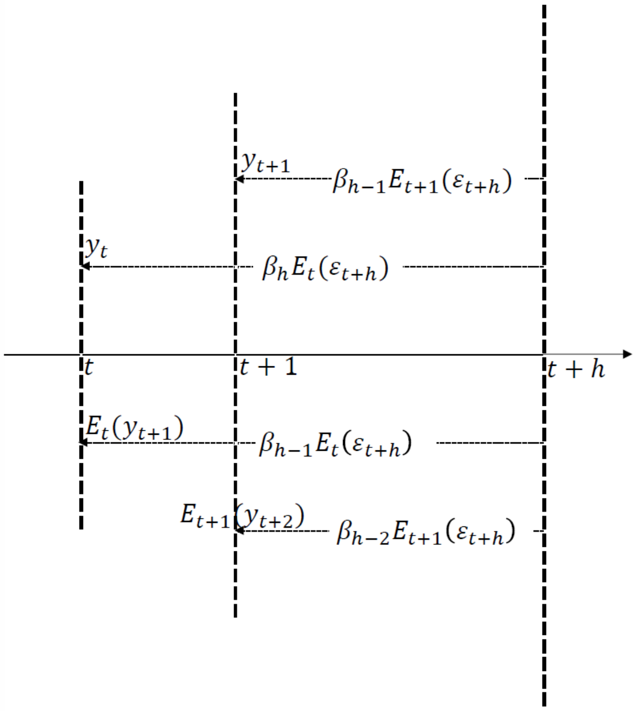
Equation (3) at time t and $t+1$ and take conditional expectation w.r.t. information set $\mathcal{I}_t$ to get

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_r y_{t-r} + \sum_{j=0}^{\infty} \beta_j E_t(\varepsilon_{t+j}) \tag{14}$$

$$E_t(y_{t+1}) = \phi_1 y_t + \cdots + \phi_r y_{t-r+1} + \sum_{j=0}^{\infty} \beta_j E_t(\varepsilon_{t+1+j}). \tag{15}$$

Note that the equations (14) and (15) are purely non-causal to provide a convenient example. Should a time series show signs of non-causality then the parameter $\beta_j \neq 0$ $(j \geq 0)$. Another difference between causal and non-causal models is that in the latter case $E_t(\varepsilon_t) \neq \varepsilon_t$ because ε_t depends on y_{t+j} $(0 < j \leq s)$, this can be seen in Equation (1).

Figure 3: **Purely Non-causal Expectation Dynamics**



y_{t+1}
$\beta_{h-1}E_{t+1}(\varepsilon_{t+h})$
y_t
$\beta_h E_t(\varepsilon_{t+h})$
t
$t+1$
$t+h$
$E_t(y_{t+1})$
$\beta_{h-1}E_t(\varepsilon_{t+h})$
$E_{t+1}(y_{t+2})$
$\beta_{h-2}E_{t+1}(\varepsilon_{t+h})$

Graphical overview of the effects of future ε on realized and expected future data, above and below timeline respectively.

The dependency structure of ε_t is more complicated in the non-causal case than in the causal case. For example an AR(0,3) model involves terms y_t up to y_{t+3}, which in turn contain respective error terms ε_t up to ε_{t+3}. Therefore to calculate y_t it is necessary to take the conditional expectation at time t of these error terms and discount the values back to time t. Consider again the AR(0,3) model situation, when calculating y_{t-1} and all information is known up to and including time t. In a causal context every term included in y_{t-1} would have been observed and no uncertainty would remain. However in a non-causal context, in this case AR(0,3), there are errors included that are observed after time t, namely ε_{t+1} and ε_{t+2}. These errors will be included in the following form: $\beta_2 E_t(\varepsilon_{t+1})$ and $\beta_3 E_t(\varepsilon_{t+2})$. A practical overview of this type of situation is summarized in Figure 3.

4.2 AR(r,s) Forecasting

To forecast future values of y_t using the AR(r,s) model it is important to note that future values of v_t are equivalent to y_t. Section 3.1 shows the equivalence of datasets $(y_1, ..., y_T)$ and $(u_1, ..., u_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, v_{T-s+1}, ..., v_T)$. If the dataset is increased to $(y_1, ..., y_{T+H})$ the equivalent set becomes $(u_1, ..., u_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, v_{T-s+1}, ..., v_{T+H})$. Equation (3) makes this equivalence possible and therefore it is feasible to find the following conditional p.d.f. for v_t

$$\begin{aligned} &l(v_{T+1}, ..., v_{T+H}|y_1, ..., y_T) \\ &= l(v_{T+1}, ..., v_{T+H}|u_1, ..., u_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}, v_{T-s+1}, ..., v_T) \\ &= l(v_{T+1}, ..., v_{T+H}|v_{T-s+1}, ..., v_T), \end{aligned} \tag{16}$$

where $u_1, ..., u_r, \varepsilon_{r+1}, ..., \varepsilon_{T-s}$ can be discarded due to independence with $v_{T-s+1}, ..., v_{T+H}$ (section 3.1). Rewrite the conditional p.d.f. (16) as

$$\begin{aligned} &l(v_{T+1}, ..., v_{T+H}|v_{T-s+1}, ..., v_T) \\ &= \frac{l(v_{T-s+1}, ..., v_T, v_{T+1}, ..., v_{T+H})}{l_s(v_{T-s+1}, ..., v_T)} \\ &= \frac{l(v_{T-s+1}, ..., v_{T+H-s}|v_{T+H-s+1}, ..., v_{T+H})}{l_s(v_{T-s+1}, ..., v_T)} l_s(v_{T+H-s+1}, ..., v_{T+H}), \end{aligned} \tag{17}$$

where l_s denotes the stationary density of s consecutive values of v. These values are denoted as $v_{\tau-s+1}, ..., v_\tau$. Note that both densities $l(\cdot)$ and $l_s(\cdot)$ are unknown. The next step is therefore to find closed-form solutions. rewriting $l(\cdot)$ gives

$$\begin{aligned} &l(v_{T-s+1}, ..., v_{T+H-s}|v_{T+H-s+1}, ..., v_{T+H}) \\ &= l(v_{T-s+1}|v_{T-s+2}, ..., v_{T+1}) l(v_{T-s+2}|v_{T-s+3}, ..., v_{T+2}) ... l(v_{T+H-s}|v_{T+H-s+1}, ..., v_{T+H}) \\ &= g(\Phi(L^{-1})v_{T-s+1}) g(\Phi(L^{-1})v_{T-s+2}) ... g(\Phi(L^{-1})v_{T+H-s}). \end{aligned} \tag{18}$$

Parametrizing the p.d.f. g by parameter θ allows the parameters Φ, Ψ and θ to be estimated using maximum likelihood. Recall that this research uses a t-distribution with ν degrees of freedom (section 3.1). For testing purposes $\nu \in \{1, 2, 3, 5, 10\}$, to see the effects of ranging from a high kurtosis with $\nu = 1$, also known as a Cauchy distribution, to a distribution that is close to Gaussian with $\nu = 10$.

The final step is to find a closed-form solution to the p.d.f. $l_s(\cdot)$. By using the joint density of $\Upsilon_{\tau-s+1}, ..., \Upsilon_\tau$ and any sequence of lagged values $\upsilon^*_{\tau-s+1}, ..., \upsilon^*_\tau$ at any date τ, use the Iterated Expectations Theorem to rewrite the stationary density.

$$\begin{aligned}
&l_s(\upsilon^*_{\tau-s+1}, ..., \upsilon^*_\tau)\\
&= E[l(\upsilon^*_{\tau-s+1}, ..., \upsilon^*_\tau | \Upsilon_{\tau+1}, ..., \Upsilon_{\tau+s})]\\
&= E[g(\upsilon^*_{\tau-s+1} - \phi_1 \upsilon^*_{\tau-s+2} - ... - \phi_s \Upsilon_{\tau+1})...g(\upsilon^*_\tau - \phi_1 \Upsilon_{\tau+1} - ... - \phi_s \Upsilon_{\tau+s})]\\
&\approx \frac{1}{T-s+1} \sum_{t=1}^{T-s+1} \{\hat{g}(\upsilon^*_{\tau-s+1} - \hat{\phi}_1 \upsilon^*_{\tau-s+2} - ... - \hat{\phi}_s \hat{\upsilon}_t)...\hat{g}(\upsilon^*_\tau - \hat{\phi}_1 \hat{\upsilon}_t - ... - \hat{\phi}_s \hat{\upsilon}_{t+s-1})\},
\end{aligned} \tag{19}$$

where the last step is the sample-based approximation of the expectation with g and ϕ replaced by their estimated values and all current and past υ replaced by their filtered values. Rewrite Equation (17) by using (18) and (19) into a closed form solution for the predictive density $\hat{\Pi}$ where $\hat{l}_s$ is evaluated at two starting dates τ, namely $\tau = T - s + 1$ and $\tau = T + H - s + 1$. For example when $s = 1$ and $H = 2$ the predictive density is

$$\hat{\Pi}(\upsilon_{T+1}, \upsilon_{T+2} | \hat{\upsilon}_T) = \frac{\hat{g}(\hat{\upsilon}_T - \hat{\phi}_1 \upsilon_{T+1}) \hat{g}(\upsilon_{T+1} - \hat{\phi}_1 \upsilon_{T+2}) \sum_{t+1}^{T} \hat{g}(\upsilon_{T+2} - \hat{\phi}_1 \hat{\upsilon}_t)}{\sum_{t=1}^{T} \hat{g}(\hat{\upsilon}_T - \hat{\phi}_1 \hat{\upsilon}_t)}.$$

For $s = 2$ and $H = 2$ the predictive density becomes

$$\begin{aligned}
\hat{\Pi}(\upsilon_{T+1}, \upsilon_{T+2} | \hat{\upsilon}_T, \hat{\upsilon}_{T-1}) = {} & \hat{g}(\hat{\upsilon}_{T-1} - \hat{\phi}_1 \hat{\upsilon}_T - \hat{\phi}_2 \upsilon_{T+1}) \hat{g}(\hat{\upsilon}_T - \hat{\phi}_1 \upsilon_{T+1} - \hat{\phi}_2 \upsilon_{T+2})\\
& \times \sum_{t+1}^{T-1} \hat{g}(\upsilon_{T+1} - \hat{\phi}_1 \upsilon_{T+2} - \hat{\phi}_2 \hat{\upsilon}_t) \hat{g}(\upsilon_{T+2} - \hat{\phi}_1 \hat{\upsilon}_t - \hat{\phi}_2 \hat{\upsilon}_{t+1})\\
& / \sum_{t=1}^{T-1} \hat{g}(\hat{\upsilon}_{T-1} - \hat{\phi}_1 \hat{\upsilon}_T - \hat{\phi}_2 \hat{\upsilon}_t) \hat{g}(\hat{\upsilon}_T - \hat{\phi}_1 \hat{\upsilon}_t - \hat{\phi}_2 \hat{\upsilon}_{t+1}).
\end{aligned}$$

Note that the fraction in front of the summation sign in Equation (19) cancels out. The predictive density provides the possibility to find future values of υ. A Sampling Importance Resampling (SIR) method, introduced by Rubin (1987), simulates values from this predictive density. The goal of the SIR method is to simulate values from a known, but non-standard, density function f. This is done by resampling simulated values of a known, simpler distribution g according to appropriate weights. In this research I choose the same distribution for g as the error term of the AR(r,s) model, a t-distribution with ν degrees of freedom. First, simulate S values X^s of the t-distribution. Second, calculate the weight $\frac{f(X^s)}{g(X^s)}$ for every s. Third, draw H (forecast horizon) values from the set X where every X^s has its corresponding weight, i.e. a higher weight causes the value more likely to be drawn.

At this point I am able to make a step-by-step forecasting procedure. First, estimate an AR(r,s) model on the data $y_1, ..., y_T$ to obtain model parameters Φ and Ψ. Second, from this data infer ε, u and υ. Third, use the predictive density in the SIR method to simulate future values of υ. Fourth, with the newly simulated values compute corresponding ε, u and y. Fifth, repeat previous steps with a dataset that is increased by H observations. Algorithm 1 describes these steps in pseudo-code for convenience.

Algorithm 1 $AR(r,s)$ model forecasting

```
while $T < T'$ do                                          ▷ $T'$ is end of forecasting period
    Infer $\hat{\varepsilon}_{r+1},...,\hat{\varepsilon}_{T-s}$, $\hat{u}_1,...,\hat{u}_{T-s}$ and $\hat{\upsilon}_{r+1},...,\hat{\upsilon}_T$ from $y_1,...,y_T$
    Estimate predictive density $\hat{\Pi}$
    Simulate $\upsilon_{T+1},...,\upsilon_{T+H}$ with SIR method
    Compute $\hat{\varepsilon}_{T-s+1},...,\hat{\varepsilon}_{T-s+H}$, $\hat{u}_{T-s+1},...,\hat{u}_{T-s+H}$ and $y_{T+1},...,y_{T+H}$
    Increase $T$ with $H$
end while
```

This algorithm describes the forecasting procedure of the AR(r,s) model with pseudo-code.

4.3 AR(p) Forecasting

The forecasting procedure of the AR(p) model is more straightforward than that of the AR(r,s) model. The main reason for this is of course the causality of AR(p) as opposed to the non-causality of AR(r,s). Recall Equation (9) and extract forecasted values by taking the conditional expectation.

$$\begin{aligned}
E_t(y_{t+1}) &= E_t(c + \sum_{i=1}^{p} \phi_i y_{t+1-i} + \varepsilon_{t+1}) \\
E_t(y_{t+1}) &= c + \sum_{i=1}^{p} \phi_i y_{t+1-i} + E_t(\varepsilon_{t+1}) \\
E_t(y_{t+1}) &\equiv \hat{y}_{t+1} \\
&= c + \sum_{i=1}^{p} \phi_i y_{t+1-i} \qquad (20)
\end{aligned}$$

where $E_t(\varepsilon_{t+1}) = 0$. After the estimation of parameter ϕ_i all elements of the expression are known. When the forecast horizon H increases, the expression will contain values y_{t+H-i} that are not in the conditional information set $\mathcal{I}_t$. This problem can be solved by recursively putting in the closed-form expression of the forecast $H-1$ months ahead. To illustrate, obtain Equation (21) by setting $H = 2$ and use (20) to simplify the expression

$$\begin{aligned}
E_t(y_{t+2}) &\equiv \hat{y}_{t+2} \\
&= c + \phi_1 \hat{y}_{t+1} + \sum_{i=2}^{p} \phi_i y_{t+2-i}. \qquad (21)
\end{aligned}$$

4.4 VAR Forecasting

Forecasting a VAR(p) model is similar to forecasting an AR(p) model. The main difference is that all elements contained in $\hat{y}_{t+H}$ are vectors and matrices instead of scalars. The VAR model is exactly reproduced from Kilian (2009), therefore the resulting time series represents the log real price of oil. The resulting time series from the AR(r,s) model represents the growth rate of the real price of oil. To compare these time series, transform the log real price of oil by taking the exponential values and then taking the yearly growth rates.

Use Equation (10) and extract forecasted values by taking the conditional expectation

$$\begin{aligned}
E_t(y_{t+1}) &= E_t(c + A_1 y_t + A_2 y_{t-1} + \cdots + A_p y_{t-p+1} + \varepsilon_{t+1}) \\
E_t(y_{t+1}) &= c + A_1 y_t + A_2 y_{t-1} + \cdots + A_p y_{t-p+1} + E_t(\varepsilon_{t+1}) \\
E_t(y_{t+1}) &\equiv \hat{y}_{t+1} \\
&= c + A_1 y_t + A_2 y_{t-1} + \cdots + A_p y_{t-p+1}
\end{aligned}$$

Where $E_t(\varepsilon_{t+1}) = 0$. Increasing the forecast horizon has similar implications as the AR(p) model. Recursively putting in the $H-1$ months ahead forecast results in

$$\begin{aligned}
E_t(y_{t+2}) &\equiv \hat{y}_{t+2} \\
&= c + A_1 \hat{y}_{t+1} + A_2 y_t + \cdots + A_p y_{t-p+2}.
\end{aligned} \tag{22}$$

4.5 Local Level Model Forecasting

To forecast values with the Local Level model I will derive the Kalman filter (1960). For convenience equations (23) and (24) denote the LLM again

$$y_t = \mu_t + \varepsilon_t \tag{23}$$

$$\mu_{t+1} = \mu_t + \eta_t. \tag{24}$$

To create the Kalman filter let $\mu_{t|j} = E(\mu_t | \mathcal{I}_j)$ be the conditional expectation of μ_t given information set $\mathcal{I}_j$ and let $\Sigma_{t|j} = Var(\mu_t | \mathcal{I}_j)$ be the conditional variance of μ_t given $\mathcal{I}_j$. $y_{t|j} = E(y_t | \mathcal{I}_j)$ denotes the conditional mean of y_t given $\mathcal{I}_j$. Furthermore let $a_t = y_t - y_{t|t-1}$ and $V_t = Var(a_t | \mathcal{I}_{t-1}) = Var(a_t)$ be one-step-ahead forecast error and its variance of y_t given $\mathcal{I}_{t-1}$. A key insight in forecasting with the Kalman filter is recognizing that it is suffices to forecast the future state μ. Therefore derive the following equality

$$y_{t|t-1} = E(y_t | \mathcal{I}_{t-1}) = E(\mu_t + \varepsilon_t | \mathcal{I}_{t-1}) = E(\mu_t | \mathcal{I}_{t-1}) = \mu_{t|t-1},$$

which implies

$$\begin{aligned}
a_t &= y_t - y_{t|t-1} = y_t - \mu_{t|t-1} \\
V_t &= Var(y_t - \mu_{t|t-1}|\mathcal{I}_{t-1}) = Var(\mu_t + \varepsilon_t - \mu_{t|t-1}|\mathcal{I}_{t-1}) \\
&= Var(\mu_t - \mu_{t|t-1}|\mathcal{I}_{t-1} + Var(\varepsilon_t|\mathcal{I}_{t-1}) = \Sigma_{t|t-1} + \sigma_\varepsilon^2.
\end{aligned}$$

Note that a_t is independent with y_j for $j < t$

$$\begin{aligned}
E(a_t) &= E[E(a_t|\mathcal{I}_{t-1})] = E[E(y_t - y_{t|t-1}|\mathcal{I}_{t-1})] = E[y_{t_t-1} - y_{t|t-1}] = 0 \\
Cov(a_t, y_j) &= E(a_t y_j) = E[E(a_t y_j|\mathcal{I}_{t-1})] = E[y_j E(a_t|\mathcal{I}_{t-1})] = 0, \quad j < t.
\end{aligned}$$

The following equality now holds for the information set, $\mathcal{I}_t = \{\mathcal{I}_{t-1}, y_t\} = \{\mathcal{I}_{t-1}, a_t\}$. Make use of the following theorem to create the Kalman filter. Suppose that $\mathbf{x}$ and $\mathbf{y}$ are random vectors such that their joint distribution is multivariate normal with mean $E(\mathbf{w}) = \boldsymbol{\mu}_w$ and covariance matrix $\boldsymbol{\Sigma}_{mw} = Cov(\mathbf{m}, \mathbf{w})$, where w and m are x and y. In addition assume that the diagonal block covariance matrix $\boldsymbol{\Sigma}_{ww}$ is non-singular for $w = x, y$. Then,

$$\begin{aligned}
E(\mathbf{x}|\mathbf{y}) &= \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y) \\
Var(\mathbf{x}|\mathbf{y}) &= \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xx}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}.
\end{aligned} \tag{25}$$

Use the preceding findings to derive the filtering and forecasting steps of the Kalman filter. Filtering is done by finding y_t given $\mathcal{I}_t$, which is equivalent to $(\mu_t, a_t)'$ given $\mathcal{I}_{t-1}$. First derive the properties of this joint distribution and then use (25) to create the filtering step. To create the joint distribution, the following expressions are needed

$$\begin{aligned}
E(a_t|\mathcal{I}_{t-1}) &= E(y_t - y_{t|t-1}|\mathcal{I}_{t-1}) = 0 \\
Var(a_t|\mathcal{I}_{t-1}) &= V_t = Var(y_t - y_{t|t-1}|\mathcal{I}_{t-1}) = Var(\mu_t + \varepsilon_t - \mu_{t|t-1}|\mathcal{I}_{t-1}) \\
&= Var(\mu_t - \mu)t|t-1|\mathcal{I}_{t-1}) + Var(\varepsilon_t|\mathcal{I}_{t-1}) = \Sigma_{t|t-1} + \sigma_\varepsilon^2 \\
E(\mu_t|\mathcal{I}_{t-1}) &= \mu_{t|t-1} \\
Var(\mu_t|\mathcal{I}_{t-1}) &= \Sigma_{t|t-1} \\
Cov(\mu_t, a_t|\mathcal{I}_{t-1}) &= E(\mu_t a_t|\mathcal{I}_{t-1}) = E(\mu_t(y_t - \mu_{t|t-1})|\mathcal{I}_{t-1}) \\
&= E(\mu_t(\mu_t + \varepsilon_t - \mu_{t|t-1})|\mathcal{I}_{t-1}) \\
&= E(\mu_t(\mu_t - \mu_{t|t-1})|\mathcal{I}_{t-1}) + E(\mu_t\varepsilon_t|\mathcal{I}_{t-1}) \\
&= E((\mu_t - \mu_{t|t-1})^2|\mathcal{I}_{t-1}) = Var(a_t|\mathcal{I}_{t-1}) = \Sigma_{t|t-1}.
\end{aligned} \tag{26}$$

These five expressions in (26) combined give the joint distribution of $(\mu_t, a_t)'$ given $\mathcal{I}_{t-1}$

shown in Equation (27)

$$\begin{bmatrix} \mu_t \\ a_t \end{bmatrix}_{\mathcal{I}_{t-1}} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{t|t-1} \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{t|t-1} & \Sigma_{t|t-1} \\ \Sigma_{t|t-1} & \Sigma_{t|t-1} + \sigma_\varepsilon^2 \end{bmatrix} \right). \tag{27}$$

Now the theorem in (25) can be applied to obtain the following expressions

$$\begin{aligned} \mu_{t|t} &= \mu_{t|t-1} + \frac{\Sigma_{t|t-1}}{\Sigma_{t|t-1} + \sigma_\varepsilon^2} a_t \\ \Sigma_{t|t} &= \Sigma_{t|t-1} - \frac{\Sigma_{t|t-1}}{\Sigma_{t|t-1} + \sigma_\varepsilon^2}. \end{aligned}$$

Finally find an expression for forecasting μ_{t+1} given $\mathcal{I}_t$

$$\begin{aligned} \mu_{t+1|t} &= E(\mu_t + \eta_t|\mathcal{I}_t) = E(\mu_t|\mathcal{I}_t) = \mu_{t|t} \\ \Sigma_{t+1|t} &= Var(\mu_{t+1}|\mathcal{I}_t) = Var(\mu_t|\mathcal{I}_t) + Var(\eta_t) = \Sigma_{t|t} + \sigma_\eta^2. \end{aligned}$$

Equation (28) provides the complete Kalman filter

$$\begin{aligned} a_t &= y_t - \mu_{t|t-1} \\ V_t &= \Sigma_{t|t-1} + \sigma_\varepsilon^2 \\ K_t &= \Sigma_{t|t-1}/V_t \\ \mu_{t+1|t} &= \mu_{t|t-1} + K_t a_t \\ \Sigma_{t+1|t} &= \Sigma_{t|t-1}(1 - K_t) + \sigma_\eta^2. \end{aligned} \tag{28}$$

To forecast values using the Kalman filter, estimate the parameters with maximum likelihood and use $y_{t+1|t} = \mu_{t+1|1} = \mu_{t|t-1} + K_t a_t$.

5 Results

This sections measures and compares quality of the forecasts made by the models discussed in section 3 with the forecasting methods discussed in section 4. The main focus of this research remains the AR(r,s) model, therefore a selection of combinations has been made regarding the lags r and leads s. This selection results in three purely non-causal combinations, namely $s = \{1, 2, 3\}$ (in combination with $r = 0$). Following the reasoning of Lanne & Saikkonen (2011) a combination of r and s that satisfies $r+s = p$, where p is the order of the best fitting AR(p) model, is suitable. Therefore I also include an AR(1,1) model, the models that satisfy this rule are AR(0,1) for electricity and gas and both AR(0,2) and AR(1,1) for oil. Note that the combinations chosen here do not represent an exhaustive set, however a selection has to be made in order to keep the amount of results manageable. The following subsections use statistics to measure and compare the forecasting performance of the different models. Evaluate the quality of a forecast by looking at the forecast error which is defined as

$$e_t = y_t - \hat{y}_t,$$

where y_t is the actual value of the time series at time t and $\hat{y}_t$ is its forecasted value. Table 1 provides the forecast period. The forecast error will therefore be available for this period as well.

5.1 Root Mean Square Forecast Error

The first forecast quality measure is the root mean square forecast error (RMSFE). This is a popular measure that returns a single value of average error. A distinct feature is the higher penalty for bigger errors, this higher penalty is a result of first squaring and then averaging the error values. Afterwards take the square root to ensure the result is interpretable. Calculate the RMSFE as follows

$$RMSFE = \sqrt{\frac{\sum_{t=1}^{n}(e_t)^2}{n}}.$$

A final important note about the RMSFE is the assumption of a finite variance, this causes the results with 1 and 2 d.o.f. in the t-distribution to be infeasible for the calculation of the RMSFE since a second moment is not defined for these cases. Table 5 contains a collection of RMSFE values. The table contains fifteen columns of values in pairs of five for each dataset. The five columns in each dataset represent a certain forecast horizon. The different models are separated with horizontal lines and every non-causal model contains RMSFE values for different degrees of freedom (d.o.f.).

Overall the RMSFE values of the AR(r,s) models are slightly greater than those of the causal models. The difference between the non-causal models is very small. The models that satisfy $r + s = p$, as mentioned before, do not seem to outperform the other models. In the causal models I expect an increasing RMSFE value when the forecast horizon increases. With the oil data, this pattern is indeed showing. However, with the electricity and gas data the values are rather constant. A reason for this could be the relatively small dataset on electricity and gas. With a forecast horizon of twelve months the model is only estimated four times. Looking again to the oil data results, the non-causal models do not show this pattern of increasing error. A reason for this behavior should be found in the difference between the models, causal models start to use data that is increasingly further removed from the forecasted date when the horizon increases. This could lead to a forecast that drifts off to a wrong direction, which of course increases the forecast error. Non-causal models on the other hand make use of a predictive density. Only one predictive density is estimated for the entire forecast horizon, but values are drawn every month. The 'old' data is therefore indirectly being used to create the forecast. This could dampen the aging effect of the data.

Table 5: **Root Mean Square Forecast Error**

		Electricity					Gas					Oil				
	Horizon →	1	2	3	6	12	1	2	3	6	12	1	2	3	6	12
Causal	AR(p)	1.2	1.2	1.2	1.2	1.1	1.7	1.7	1.7	1.7	1.8	7.2	7.2	7.6	7.8	7.8
	LLM	1.1	1.1	1.1	1.1	1.1	5.1	6.7	1.6	1.7	1.7	7.2	7.3	7.6	7.8	7.8
AR(0,1)	3 d.o.f.	1.6	1.6	1.4	1.7	1.5	1.9	1.8	1.9	2.1	2.0	10.1	11.2	12.4	11.2	10.0
	5 d.o.f.	1.3	1.3	1.5	1.5	1.3	2.0	2.0	1.7	1.8	2.0	10.6	11.2	11.4	10.7	10.6
	10 d.o.f.	1.4	1.2	1.5	1.5	1.2	2.2	2.3	2.0	2.1	2.0	11.3	11.6	12.5	12.9	11.8
AR(0,2)	3 d.o.f.	1.4	1.6	1.2	1.3	1.5	1.9	1.8	1.7	1.7	1.7	11.1	12.0	11.5	12.2	11.8
	5 d.o.f.	1.3	1.5	1.4	1.3	1.6	2.0	1.9	1.7	1.9	1.7	12.1	11.1	12.2	12.3	13.0
	10 d.o.f.	1.4	1.3	1.3	1.5	1.4	1.9	1.8	2.0	2.0	1.6	12.8	13.4	13.5	13.7	13.5
AR(0,3)	3 d.o.f.	1.5	1.3	1.1	1.7	1.5	1.7	1.9	1.6	1.6	1.6	10.9	11.2	12.1	12.3	10.6
	5 d.o.f.	1.9	1.4	1.6	1.4	1.5	1.8	1.7	2.0	1.7	1.8	12.3	12.6	12.5	11.8	11.9
	10 d.o.f.	1.4	1.5	1.5	1.3	1.4	2.1	1.9	1.8	1.8	2.2	13.0	12.7	13.7	13.4	13.7
AR(1,1)	3 d.o.f.	1.5	1.6	1.3	1.3	1.5	2.0	1.7	2.0	1.8	2.1	11.0	12.3	12.2	10.6	14.4
	5 d.o.f.	1.3	1.4	1.6	1.1	1.5	2.0	1.8	2.0	1.8	1.9	10.5	10.6	11.6	13.4	10.8
	10 d.o.f.	1.4	1.2	1.3	1.4	1.4	2.1	2.3	2.2	2.2	2.1	11.4	12.0	13.2	13.5	14.6

All values represent root mean square forecast error (RMSFE) values. For every dataset the values are presented for every forecast horizon (horizontal) and degree of freedom in the t-distribution of the error term ε_t in the AR(r,s) model (vertical).

5.2 Mean Absolute Forecast Error

The second forecast quality measure is the mean absolute forecast error (MAFE). This measure returns a single value of average error, just like the RMSFE. The penalty for bigger errors is less severe, since the errors are not squared. Calculate the MAFE as follows

$$MAFE = \frac{\sum_{t=1}^{n} |e_t|}{n}. \tag{29}$$

No assumption has to be made regarding a finite variance, therefore the results of 1 and 2 d.o.f. can be calculated as opposed to the RMSFE. Table 6 shows the MAFE and has a similar layout as table 5.

All values in Table 6 are smaller than their counterparts in Table 5. This is the result of the lower penalty of error magnitude. All other patterns that have been discussed regarding the RMSFE hold for the MAFE as well. Looking at the 1 and 2 d.o.f. values we see no notable results at 2 d.o.f., however the 1 d.o.f. MAFE values are regularly greater than the other degrees of freedom. 1 d.o.f. implies a fat-tailed t-distribution which seems to be worse than the distributions with (slightly) less fat-tails.

Table 6: **Mean Absolute Error**

		Electricity					Gas					Oil				
	Horizon →	1	2	3	6	12	1	2	3	6	12	1	2	3	6	12
Causal	AR(p)	0.9	0.9	0.9	0.8	0.7	1.1	1.1	1.1	1.1	1.2	5.7	5.6	5.8	5.8	5.8
	LLM	0.8	0.8	0.8	0.8	0.7	1.7	2.3	1.0	1.0	1.1	5.7	5.6	5.8	5.9	5.8
AR(0,1)	1 d.o.f.	1.4	1.2	1.6	3.1	1.2	1.2	1.4	2.0	2.0	1.3	8.1	8.5	8.8	8.0	7.9
	2 d.o.f.	1.0	1.3	1.1	1.1	1.0	1.0	1.2	1.6	1.2	1.4	7.8	7.7	8.6	8.2	8.0
	3 d.o.f.	1.0	1.1	1.0	1.2	1.0	1.2	1.2	1.3	1.4	1.3	7.8	8.9	9.3	8.3	7.7
	5 d.o.f.	1.0	1.0	1.0	1.1	1.1	1.4	1.3	1.2	1.3	1.4	8.6	9.2	9.1	8.4	8.3
	10 d.o.f.	1.1	0.9	1.2	1.1	1.0	1.4	1.8	1.4	1.5	1.6	9.5	9.7	10	10.6	9.4
AR(0,2)	1 d.o.f.	1.7	1.2	1.7	1.3	1.7	2.1	1.8	1.4	1.2	1.4	8.5	9.0	8.1	9.2	7.3
	2 d.o.f.	1.0	1.0	1.2	0.9	1.2	1.2	1.4	1.5	1.4	1.2	9.1	8.6	9.2	8.6	8.6
	3 d.o.f.	1.0	1.1	0.9	1.0	1.0	1.3	1.2	1.2	1.1	1.2	8.9	9.3	9.1	9.6	9.4
	5 d.o.f.	1.0	1.1	1.0	0.9	1.2	1.4	1.3	1.2	1.2	1.2	9.8	8.9	10.1	9.7	10.6
	10 d.o.f.	1.1	0.9	0.9	1.1	1.0	1.4	1.3	1.4	1.5	1.1	10.8	10.8	11.1	11.7	11.4
AR(0,3)	1 d.o.f.	1.1	1.2	1.0	1.6	1.3	1.4	1.3	1.4	1.2	1.2	8.5	8.3	8.6	9.8	7.8
	2 d.o.f.	1.1	1.3	1.0	1.0	1.0	1.1	1.1	1.4	1.2	1.1	8.2	8.6	9	8.7	8.4
	3 d.o.f.	1.0	0.9	0.8	1.1	1.1	1.1	1.3	1.0	1.2	1.1	8.7	9.0	9.4	9.8	8.4
	5 d.o.f.	1.3	1.0	1.1	1.1	1.2	1.3	1.1	1.3	1.1	1.3	10.1	10.1	10	9.5	9.7
	10 d.o.f.	1.1	1.1	1.1	1.0	1.1	1.6	1.3	1.3	1.3	1.6	11.0	10.4	11.6	10.9	11.6
AR(1,1)	1 d.o.f.	1.5	1.3	1.6	1.6	1.4	2.4	1.3	1.6	1.6	1.5	10.1	10.9	10.9	11.1	10.7
	2 d.o.f.	1.1	1.0	1.0	1.2	1.3	1.4	1.2	1.1	1.2	1.3	9.4	8.9	9.1	8.4	10.5
	3 d.o.f.	1.1	1.2	1.0	0.9	1.0	1.4	1.1	1.2	1.4	1.4	8.6	9.0	9.0	8.2	11.3
	5 d.o.f.	0.9	1.1	1.2	0.9	1.1	1.5	1.2	1.3	1.2	1.4	8.4	8.6	9.4	10.5	8.5
	10 d.o.f.	1.0	1.0	1.0	1.1	1.1	1.5	1.6	1.6	1.6	1.6	9.2	10.0	11.1	11.5	12.6

All values represent mean absolute forecast error (MAE) values. For every dataset the values are presented for every forecast horizon (horizontal) and degree of freedom in the t-distribution of the error term ε_t in the AR(r,s) model (vertical).

5.3 Diebold Mariano Test

Diebold and Mariano (1995) propose a test to compare the quality of different forecasts. The null hypothesis of the test is $E[d_t] = 0$ with a two-sided alternative hypothesis, where d_t is the difference in loss functions $g(e_{it})$. The loss function can take on any form, I choose a quadratic form in this research to increasingly penalize greater errors, $g(e_{it}) = e_{it}^2, \quad i = \{1, 2\}$. The test statistic is

$$\frac{\bar{d}}{\sqrt{\hat{\sigma}_{d_t}^2 / T}} \sim \mathcal{N}(0,1), \tag{30}$$

where $\hat{\sigma}_{d_t}^2$ is the variance of d_t. Compute $\hat{\sigma}_{d_t}^2$ as follows

$$\hat{\sigma}_{d_t}^2 = \hat{\gamma}_0 + 2 \sum_{j=1}^{h-1} \hat{\gamma}_j,$$

with $\hat{\gamma}_j$ denoting the j-th order sample autocovariance

$$\hat{\gamma}_j = \frac{1}{T} \sum_{t=|j|+1}^{T} (d_t - \bar{d})(d_{t-|j|} - \bar{d}).$$

The correction of the sample variance $\hat{\gamma}_0$ with the autocovariances $\hat{\gamma}_j, j = 1, \ldots, h-1$, is based on the fact that forecast errors for h-step-ahead forecasts are serially correlated up to (at least) order $h-1$ by construction. Diebold and Mariano (1995) point out that the estimate $\hat{\sigma}_{d_t}^2$ can, in rare cases, be negative. In this case $\hat{\sigma}_{d_t}^2$ is treated as 0 and the null hypothesis of equal forecast accuracy is immediately rejected (p. 254).

Table 7 shows the test statistics. In every test e_{1t} is the forecast error of the AR(r,s) model and e_{2t} the forecast error of the model of interest. The standard normal distributed test statistic therefore indicates that there is reason to believe the forecast quality of the AR(r,s) model is better if the statistic is smaller than -1.96 and the reverse holds true when the statistic is greater than 1.96. The models included in Table 7 are those that comply with the $r + s = p$ rule. The Diebold Mariano statistics of all other models discussed in this section can be found in appendix C. The left-hand side of Table 7 gives the statistics when AR(r,s) is compared to AR(p) and the right-hand side does this for the Local Level model. Electricity and Gas results show no forecast is significantly outperforming the causal forecasts. In some cases the non-causal forecasts are significantly worse, namely when the statistic is greater than the critical value of 1.96. In most cases there is no clear 'winner'. Oil forecasts however, are almost always outperformed by the causal forecasts. The difference between different d.o.f. is clearly visible in the oil forecasts. More degrees of freedom reduce the quality of the non-causal forecast. For electricity and gas this pattern is not always visible, which

Table 7: **Diebold Mariano Statistics**

		AR(p)					LLM				
	Horizon →	1	2	3	6	12	1	2	3	6	12
Electricity AR(0,1)	1 d.o.f.	1.23	1.19	1.50	1.12	1.51	0.24	1.41	1.40	1.27	0.99
	2 d.o.f.	0.80	1.43	2.22	0.99	1.31	-1.01	0.80	2.11	0.73	1.54
	3 d.o.f.	1.15	1.57	1.64	1.87	1.49	1.39	1.28	1.02	1.69	1.32
	5 d.o.f.	0.97	0.47	1.69	2.28	1.37	2.01	2.73	-0.04	0.97	1.27
	10 d.o.f.	1.57	-0.14	2.23	1.50	1.13	2.71	3.29	1.78	1.56	1.65
Gas AR(0,1)	1 d.o.f.	1.29	1.33	1.55	1.12	1.40	-0.98	-0.96	1.44	1.27	1.19
	2 d.o.f.	1.34	1.56	2.32	1.10	1.25	-1.01	-0.99	2.33	0.81	1.06
	3 d.o.f.	1.35	2.09	1.92	1.85	1.74	-0.96	-0.99	1.45	1.72	1.22
	5 d.o.f.	1.63	1.92	1.85	2.13	1.45	-0.93	-0.97	0.75	1.02	1.64
	10 d.o.f.	2.85	0.45	2.29	1.51	1.11	-0.90	-0.94	2.02	1.48	1.21
Oil AR(0,2)	1 d.o.f.	5.16	2.24	4.25	2.31	2.68	5.12	2.24	4.21	2.30	2.66
	2 d.o.f.	1.75	4.27	4.43	2.44	2.79	1.75	4.28	4.39	2.44	2.79
	3 d.o.f.	6.96	4.84	4.39	2.87	2.85	6.90	4.85	4.37	2.86	2.85
	5 d.o.f.	8.81	6.03	4.92	2.84	2.90	8.66	5.99	4.88	2.84	2.90
	10 d.o.f.	10.56	5.55	5.22	3.97	2.92	10.40	5.49	5.20	3.96	2.92
Oil AR(1,1)	1 d.o.f.	4.79	3.46	3.11	2.61	1.46	4.77	3.45	3.12	2.6	1.46
	2 d.o.f.	4.87	3.19	3.21	3.00	2.05	4.84	3.19	3.22	2.97	2.06
	3 d.o.f.	5.82	4.12	2.65	3.02	2.38	5.73	4.12	2.65	3.02	2.39
	5 d.o.f.	7.06	5.05	4.80	3.34	2.42	6.97	5.06	4.73	3.33	2.43
	10 d.o.f.	7.59	6.12	5.49	4.02	2.93	7.43	6.14	5.48	4.02	2.93

Diebold-Mariano statistics for electricity AR(0,1), gas AR(0,1) and oil AR(0,2) & AR(1,1). A lower statistic is in favor of the AR(r,s) forecast quality. Note that the critical values are all roughly ±2.

can be attributed to the smaller sample size. The difference in quality over different forecast horizons is not visible in the electricity and gas data, but in the oil data the test statistic is decreasing when the forecast horizon increases. Keep in mind that this could simply show that at long forecast horizons the different models simply perform equally bad.

In addition to the forecast results over the entire forecast period it could be interesting to evaluate forecasting performance in subperiods. Table 8 divides the forecast period in subperiods for the oil dataset. The crisis period is defined as the period from September 2007 up to and including December 2010. The forecasts before, during and after this period are compared. The AR(1,1) model will be used because the Diebold Mariano statistics of this model are on average lower (favoring the AR(r,s) model) than the AR(0,1) model. The period during the crisis has, on average, the lowest Diebold Mariano statistics, indicating the relative best performance of the non-causal model. This in addition to the overall poor performance of the non-causal model in oil price forecasting, leads to believe that the only reason why the test statistics are lower during crisis is because the causal models simply perform worse in this period. The average value of the test statistics before the crisis are greater than after the crisis. Around the year 2003 the oil started to be more volatile which has a negative impact on forecast results.

Table 8: **Diebold Mariano Statistics Credit Crisis**

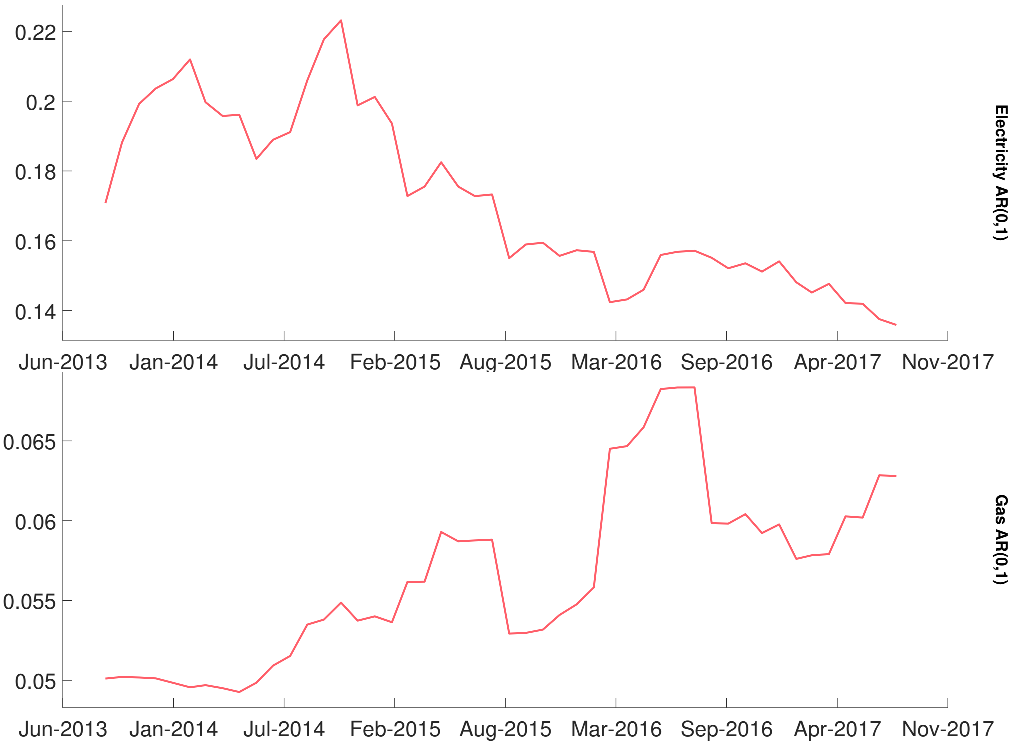
		AR(p)					LLM				
	Horizon →	1	2	3	6	12	1	2	3	6	12
Before Crisis AR(1,1)	1 d.o.f.	4.61	2.56	2.39	2.15	1.25	4.58	2.56	2.39	2.15	1.25
	2 d.o.f.	3.85	2.58	2.44	2.44	1.49	3.83	2.58	2.45	2.41	1.49
	3 d.o.f.	3.69	2.73	1.87	2.45	1.73	3.62	2.74	1.86	2.47	1.73
	5 d.o.f.	4.92	3.36	4.06	2.58	1.87	4.86	3.35	4.00	2.58	1.88
	10 d.o.f.	5.91	4.54	4.18	3.05	2.16	5.71	4.52	4.16	3.05	2.16
During Crisis AR(1,1)	1 d.o.f.	1.54	1.67	1.86	1.65	1.02	1.53	1.67	1.86	1.65	1.02
	2 d.o.f.	1.88	1.96	1.93	1.05	1.44	1.87	1.94	1.87	0.95	1.43
	3 d.o.f.	3.01	1.73	1.69	0.91	1.24	2.86	1.72	1.54	0.89	1.24
	5 d.o.f.	2.45	2.10	1.90	1.18	0.53	2.31	2.15	1.84	1.17	0.52
	10 d.o.f.	2.71	1.69	1.91	1.19	1.23	2.64	1.72	1.88	1.17	1.22
After Crisis AR(1,1)	1 d.o.f.	2.02	2.65	2.11	1.57	1.66	2.02	2.66	2.16	1.58	1.68
	2 d.o.f.	3.21	2.44	2.36	1.55	1.46	3.23	2.49	2.39	1.56	1.47
	3 d.o.f.	4.87	2.94	2.52	1.77	1.57	4.84	2.97	2.55	1.78	1.58
	5 d.o.f.	4.47	3.36	2.65	1.98	1.44	4.49	3.42	2.67	1.98	1.45
	10 d.o.f.	5.45	3.87	3.25	2.37	1.76	5.45	3.92	3.26	2.37	1.77

Diebold Mariano statistics from forecasts of AR(1,1) model versus causal models for oil data. A lower statistic is in favor of the AR(r,s) forecast quality. Crisis period is considered September 2007 up to and including December 2010.

5.4 Robustness

Figures 4 and 5 show the evolution of parameter estimates as the estimation period increases during forecasting. The figures show the AR(0,1) model for electricity and gas and the AR(0,2) model for oil respectively. All parameter estimates are based on models with 1 degree of freedom and a forecast horizon of one month. I expect to see variation in the estimates when the estimation period is small, however as this period increases the estimates should stabilize. For electricity in Figure 4 this pattern holds true. The first estimate is based on 32 datapoints and varies around 0.2. As the period increases the estimates decrease towards around 0.14 and do not jump up and down a lot. Gas in Figure 4 seems to show a somewhat different pattern, the estimates are varying even at increasingly more available datapoints. However, the vertical axis shows that the movement is between 0.05 and 0.065. The magnitude of these varying estimates is therefore small and can be attributed to the random nature of the economic time series.

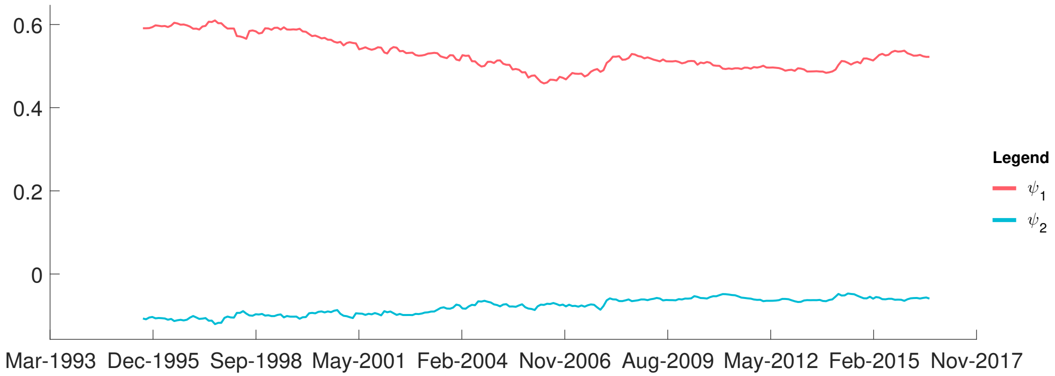
Figure 4: **parameter evolution electricity and gas**



Parameter estimates of the AR(0,1) parameter ψ for the electricity (top) and gas (bottom) datasets with a forecast horizon of one month. Vertical axis contains the estimated value and horizontal axis contains the date up to which data is used to make the estimation.

The parameter estimates for oil in Figure 5 are relatively constant throughout the entire period. This is because the oil dataset is larger than the electricity and gas datasets.

Figure 5: **parameter evolution oil**



Parameter estimates of the AR(0,2) parameters ψ_1 and ψ_2 for the oil dataset with a forecast horizon of one month. Vertical axis contains the estimated value and horizontal axis contains the date up to which data is used to make the estimation.

5.5 VAR Model Results

The VAR model creates all preceding statistics and results as well. Note that the VAR model is only used with oil data. Table 9 and Table 10 summarize the information. Table 9 presents the MAFE and RMSFE of the VAR model forecast. These values are quite constant throughout the different forecast horizons. Table 10 shows the Diebold-Mariano test statistics of the causal models versus the VAR model and the AR(r,s) model versus the VAR model. A higher statistic indicates a higher VAR forecast quality and vice versa for the other models. Similar to the AR(r,s) model, the VAR model is not able to significantly outperform the other causal models in terms of forecast quality. Moreover, the causal forecasts all perform significantly better than the VAR forecasts. Again we see that at long forecast horizons, the forecasts are starting to become equally bad and the test statistic is going towards zero. When comparing the non-causal model to the VAR model, we see test results inside the the critical values for most of the forecasts. Longer forecast horizons see to have a worse effect on the non-causal model than on the VAR model.

Table 9: **Mean Absolute Error & Root Mean Square Error VAR**

	1	2	3	6	12
MAFE	8.2	8.2	8.2	8.2	8.2
RMSFE	10.3	10.3	10.3	10.3	10.3

Mean absolute error and root mean square error values of the forecasts made by the VAR model.

Table 10: **Diebold Mariano Statistics VAR**

		1	2	3	6	12
	AR(2)	-6.91	-4.64	-4.22	-3.25	-2.76
	LLM	-6.82	-4.73	-4.38	-3.25	-2.76
AR(0,2)	1 d.o.f.	1.78	1.18	1.52	3.02	4.38
	2 d.o.f.	1.28	1.20	2.03	1.22	3.03
	3 d.o.f.	0.70	2.05	1.49	2.47	3.54
	5 d.o.f.	1.57	0.94	1.61	1.67	3.43
	10 d.o.f.	-1.25	0.60	1.73	2.32	2.34

Diebold Mariano statistics for the VAR model versus the models shown on the left-hand side of the table. A higher statistic is in favor of the VAR model.

6 Conclusion

This section includes a summary of the results and some points of discussion on the AR(r,s) model and thoughts on future research. The goal of this research is to answer the question if a non-causal process improves forecasting accuracy of economic time series compared to causal processes. In short, no it does not for the non-causal processes used in this research. However, the non-causal processes in this research do not perform significantly worse in some cases than the existing models. Namely the performance in forecasting electricity and gas data is comparable to the existing causal models.

The insights gained from the results section are 1.) Lower degrees of freedom in a t-distributed error term of the AR(r,s) model increase forecast quality relative to the causal models. This can be attributed to the higher kurtosis of economic time series in general, which is a property of a t-distribution with few d.o.f. 2.) The $r + s = p$ rule used in Lanne & Saikkonen (2011) is not supported nor rejected by the results from this research. The number of leads in a non-causal model could depend heavily on the chosen dataset and there is, as with causal models, no best number of leads to include in every situation. 3.) the AR(r,s) model shows robust parameter estimation which makes it feasible to use in practice. 4.) The use of non-causal models in forecasting economic time series is relatively new and shows a comparable forecast quality to the well-known VAR model.

The role of non-causal models in forecasting is one that needs more exploration. Decreasing forecast errors with lower degrees of freedom indicate the importance of variance. A downside of the AR(r,s) model is the approximation of the predictive density. The SIR method is computationally intensive. The randomness involved in getting the 'correct' draws from this density is an implied obstacle for any non-causal model. Future research could improve on this part of the process by using less intensive methods in which more simulations can easily be made. Furthermore future research could try and find an optimal error distribution in the AR(r,s) model. Of course, not only the setup of the AR(r,s) model could be changed but different data can also be used, such as certain stocks or financial markets.

References

Barksy, R. B., & Kilian, L. (2002). Do we really know that oil caused the great stagflation? a monetary alternative. *NBER Macroeconomics Annual*, *16*, 137-183.

Baumeister, C., & Peersman, G. (2013). The role of time-varying price elasticities in accounting for volatility changes in the crude oil market. *Journal of Applied Econometrics*, *28*(7), 1087-1109.

Berndt, E. R., Hall, B., Hall, R., & Hausman, J. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, *3*(4), 653-665.

Campbell, J. Y., Lo, A. Y., & MacKinlay, A. C. (1997). *Econometrics of financial markets.* Princeton University Press.

Davis, R., & Resnick, S. (1986). Limit theory for the sample covariance and correlation functions of moving averages. *Annals of Statistics*, *14*(2), 533-558.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366), 427-431.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253-263.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987-1008.

Engle, R. F. (2001). Garch 101: The use of ARCH/GARCH models in applied econometrics. *Journal of Economic Perspectives*, *15*(4), 157-168.

Falnes, J. (1995). On non-causal impulse response functions related to propagating water waves. *Applied Ocean Research*, *17*(6), 379-389.

Gassiat, E. (1993). Adaptive estimation in noncausal stationary AR processes. *Annals of Statistics*, *21*(4), 2022-2042.

Gourieroux, C., & Jasiak, J. (2016). Filtering, prediction and simulation methods for noncausal processes. *Journal of Time Series Analysis*, *37*(3), 405-430.

Hamilton, J. D. (1994). *Time series analysis.* Princeton University Press.

Hansen, J. V., McDonald, J. B., & Nelson, R. D. (2006). Some evidence on forecasting time-series with support vector machines. *Journal of the Operational Research Society*, *57*(9), 1053-1063.

Hansen, L. P., & Sargent, T. J. (1991). *Rational Expectations Econometrics.* Westview, Boulder, Colorado.

Huang, J., & Pawitan, Y. (2000). Quasi-likelihood estimation of non-invertible moving average processes. *Scandinavian Journal of Statistics*, *27*(4), 689-702.

Jarque, C. M., & Bera, A. K. (1980). Efficient test for normality, homoskedasticity and serial independence of regression residuals. *Economic Letters*, *6*(3), 255-259.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35-45.

Kilian, L. (2008). A comparison of the effects of exogenous oil supply shocks on output and inflation in the G7 countries. *Journal of the European Economic Association*, *6*(1), 78-121.

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), 1053-1069.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*, 159-178.

Lanne, M., & Saikkonen, P. (2011). Noncausal autoregressions for economic time series. *Journal of Time Series Econometrics*, *3*(3), 1-28.

Rosenblatt, M. (1995). Prediction and non-gaussian autoregressive stationary sequences. *Annals of Applied Probability*, *5*(1), 239-247.

Rosenblatt, M. (2000). *Gaussian and non-gaussian linear time series and random fields.* Springer-Verlag New York.

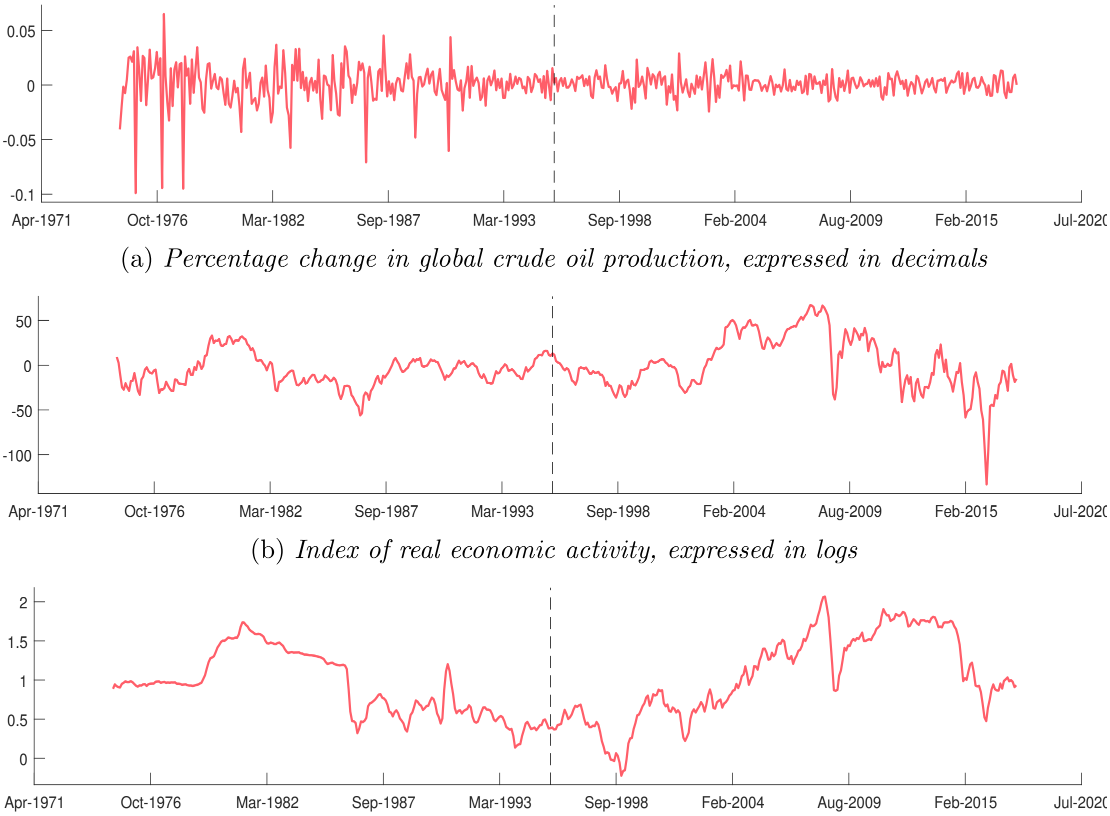
Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, *82*(398), 543-546.

Weiss, G. (1975). Time-reversibility of linear stochastic processes. *Journal of Applied Probability*, *12*(4), 831-836.

A VAR Data

Figure 6 show the data that is used in the VAR model. The data and model are exactly replicated from Kilian (2009).

Figure 6: **VAR Data**



(a) *Percentage change in global crude oil production, expressed in decimals*

(b) *Index of real economic activity, expressed in logs*

(c) *Real price of oil, expressed in logs*

B AR(p) Parameter Estimates

Table 11: **AR(p) Parameter Estimates**

Electricity						
	Constant	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	y_{t-5}
$AR(1)$	-0.23	0.35				
$AR(2)$	-0.32	0.34	-0.34			
$AR(3)$	-0.26	0.37	-0.34	0.22		
$AR(4)$	-0.22	0.36	-0.32	0.15	0.02	
$AR(5)$	-0.29	0.35	-0.30	0.07	0.01	-0.00
Gas						
	Constant	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	y_{t-5}
$AR(1)$	-0.05	0.11				
$AR(2)$	-0.05	0.10	0.07			
$AR(3)$	-0.04	0.09	0.05	0.17		
$AR(4)$	-0.04	0.08	0.05	0.17	0.03	
$AR(5)$	-0.04	0.08	0.05	0.17	0.03	0.01
Oil						
	Constant	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	y_{t-5}
$AR(1)$	0.14	0.49				
$AR(2)$	0.16	0.55	-0.13			
$AR(3)$	0.18	0.54	-0.08	-0.08		
$AR(4)$	0.19	0.54	-0.09	-0.05	-0.06	
$AR(5)$	0.19	0.53	-0.09	-0.05	-0.04	-0.04

C Diebold Mariano Statistics

Table 12: **Diebold Mariano Statistics Electricity**

		AR(p)					LLM				
	Horizon →	1	2	3	6	12	1	2	3	6	12
Electricity AR(0,2)	1 d.o.f.	1.41	1.23	1.13	1.45	1.25	1.43	1.26	1.13	1.50	1.27
	2 d.o.f.	1.57	1.22	1.72	1.21	1.53	2.92	1.61	1.96	1.52	1.44
	3 d.o.f.	1.83	1.44	-0.45	1.43	1.43	2.74	1.75	0.04	1.27	1.43
	5 d.o.f.	0.70	1.26	1.94	1.29	1.4	1.77	1.86	2.04	2.31	1.44
	10 d.o.f.	1.33	0.17	0.74	1.90	1.13	2.31	1.8	1.22	1.85	1.00
Electricity AR(0,3)	1 d.o.f.	1.37	1.25	2.09	1.47	1.19	1.42	1.4	2.22	1.49	1.19
	2 d.o.f.	1.34	1.12	1.19	1.38	1.26	1.83	1.65	1.72	1.48	1.24
	3 d.o.f.	1.45	2.11	0.67	1.27	1.40	2.29	2.65	1.34	1.29	1.34
	5 d.o.f.	0.27	1.48	2.42	-0.70	1.62	1.23	2.73	2.39	-0.31	1.71
	10 d.o.f.	1.48	-0.08	0.53	1.30	1.43	2.63	0.95	1.19	1.50	1.28
Electricity AR(1,1)	1 d.o.f.	1.03	1.70	1.00	1.57	1.29	1.11	1.91	1.10	1.57	1.28
	2 d.o.f.	1.19	1.09	1.09	0.82	2.04	1.44	1.13	1.39	0.87	-
	3 d.o.f.	1.49	0.28	-1.04	1.23	1.75	2.29	1.24	-0.54	1.28	1.77
	5 d.o.f.	1.70	1.05	1.85	1.61	1.44	1.94	1.67	2.03	1.72	1.44
	10 d.o.f.	1.63	2.05	1.65	2.03	1.60	3.32	2.78	1.77	1.82	1.56

Diebold-Mariano statistics for electricity AR(0,2), AR(0,3) and AR(1,1). A lower statistic is in favor of the AR(r,s) forecast quality. Note that the critical values are all roughly ±2. '-' is a result of a negative $\hat{\sigma}^2_{d_t}$ (section 5.3).

Table 13: **Diebold Mariano Statistics Gas**

		AR(p)					LLM				
	Horizon →	1	2	3	6	12	1	2	3	6	12
Gas AR(0,2)	1 d.o.f.	2.25	1.33	1.60	-0.38	2.20	-0.64	-0.76	1.85	-0.10	3.23
	2 d.o.f.	1.26	1.77	2.36	1.06	0.85	-0.96	-0.96	2.74	1.37	0.59
	3 d.o.f.	2.03	0.43	0.28	-0.42	-0.23	-0.95	-0.99	1.26	0.01	-0.10
	5 d.o.f.	1.79	2.11	0.21	0.96	-0.43	-0.93	-0.98	1.18	1.15	-0.03
	10 d.o.f.	1.38	0.44	2.01	1.52	-0.57	-0.95	-0.99	2.43	1.68	-0.58
Gas AR(0,3)	1 d.o.f.	1.24	1.69	1.47	1.26	0.62	-0.92	-0.97	1.57	1.34	1.57
	2 d.o.f.	0.64	1.00	1.13	0.09	-	-0.98	-1.00	1.17	0.48	-
	3 d.o.f.	0.09	1.78	-1.61	-1.08	-0.90	-0.99	-0.98	0.02	-0.38	-1.14
	5 d.o.f.	1.11	-0.34	2.53	-0.30	0.57	-0.96	-1.00	2.67	0.19	0.62
	10 d.o.f.	1.98	0.95	0.34	0.50	1.38	-0.92	-0.98	0.72	1.00	1.20
Gas AR(1,1)	1 d.o.f.	1.32	1.33	1.67	1.49	1.00	0.31	-0.96	1.74	1.58	1.14
	2 d.o.f.	2.13	0.38	0.34	0.95	0.79	-0.9	-1.00	1.30	1.08	0.98
	3 d.o.f.	1.47	-0.21	1.40	0.77	1.48	-0.93	-1.00	2.23	1.32	1.35
	5 d.o.f.	2.52	0.92	2.48	0.41	1.56	-0.94	-0.99	2.80	0.86	1.30
	10 d.o.f.	3.07	2.65	1.68	1.86	1.63	-0.92	-0.95	2.06	1.97	1.38

Diebold-Mariano statistics for gas AR(0,2), AR(0,3) and AR(1,1). A lower statistic is in favor of the AR(r,s) forecast quality. Note that the critical values are all roughly ±2. '-' is a result of a negative $\hat{\sigma}^2_{d_t}$ (section 5.3).

Table 14: **Diebold Mariano Statistics Oil**

		AR(p)					LLM				
	Horizon →	1	2	3	6	12	1	2	3	6	12
Oil AR(0,1)	1 d.o.f.	4.78	4.02	3.05	2.48	2.31	4.74	4.04	3.05	2.47	2.31
	2 d.o.f.	5.25	3.63	3.18	2.83	2.32	5.17	3.65	3.19	2.83	2.32
	3 d.o.f.	5.65	5.19	3.50	3.30	2.75	5.54	5.22	3.50	3.30	2.76
	5 d.o.f.	7.11	5.75	4.65	3.14	2.73	7.04	5.78	4.66	3.15	2.73
	10 d.o.f.	9.63	6.27	4.70	3.75	3.08	9.49	6.32	4.69	3.76	3.08
Oil AR(0,3)	1 d.o.f.	4.83	4.22	2.85	2.35	2.62	4.80	4.22	2.84	2.35	2.60
	2 d.o.f.	6.47	5.34	3.60	3.11	2.27	6.34	5.33	3.59	3.09	2.26
	3 d.o.f.	7.49	5.96	4.10	3.15	2.72	7.35	5.93	4.11	3.15	2.73
	5 d.o.f.	7.78	6.67	4.84	3.05	3.02	7.68	6.63	4.81	3.04	3.03
	10 d.o.f.	9.93	5.53	6.02	3.24	3.06	9.79	5.46	5.98	3.24	3.06

Diebold-Mariano statistics for oil AR(0,1) and AR(0,3). A lower statistic is in favor of the AR(r,s) forecast quality. Note that the critical values are all roughly ±2. '-' is a result of a negative $\hat{\sigma}^2_{d_t}$ (section 5.3).