University of Groningen

Probability density forecasts for steam coal prices in China

Ding, Lili; Zhao, Zhongchao; Han, Meng

Published in:
Energy

DOI:
[10.1016/j.energy.2021.119758](https://doi.org/10.1016/j.energy.2021.119758)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Ding, L., Zhao, Z., & Han, M. (2021). Probability density forecasts for steam coal prices in China: The role of high-frequency factors. *Energy*, 220, [119758]. <https://doi.org/10.1016/j.energy.2021.119758>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Probability density forecasts for steam coal prices in China: The role of high-frequency factors

Lili Ding ^{a, b}, Zhongchao Zhao ^a, Meng Han ^{c, *}

^a School of Economics, Ocean University of China, Qingdao, 266100, China

^b Marine Development Studies Institute of OUC, Key Research Institute in University, Ministry of Education, Qingdao, 266100, China

^c Department of Economics and Business, University of Groningen, 9747, AE, Groningen, Netherlands

ARTICLE INFO

Article history:

Received 22 April 2020

Received in revised form

29 November 2020

Accepted 30 December 2020

Available online 4 January 2021

Keywords:

Steam coal prices

High-frequency factor

MIDAS regression

XGBoost

Probability density forecast

Abstract

Coal plays a key role in China's economy as a dominant primary energy resource. In this paper, we provide probability density forecasts for weekly steam coal prices in China based on daily factors such as renewable energy source, Daqing oil, Japanese natural gas, Australia steam coal prices, coal mining industry index, A-share power sector index, A-share index, coal industry index, and temperature. The empirical results show that the influence of temperature lasts longer than other factors, while the Australia steam coal prices, renewable energy source and A-share index are the three best predictors for steam coal prices. It is also shown that the high-frequency factors are useful to forecast steam coal prices and that considering the nonlinearity of coal prices can improve the forecast accuracy by about 22%. We further provide the probability density forecasts for steam coal prices based on the influence of all the selected factors, the results suggest that our proposed method can provide accurate and satisfying probability density forecasts. Given these results, the policy-makers can make effective strategies which can not only adjust the energy structure but also ensure economic growth.

© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

In recent years, China has been facing great pressure to control greenhouse gas emissions with rapid economic growth. Power plants make a significant contribution to those emissions [1]. In spite of the promotion of sustainable development and renewable energy, coal-fired power plants still generate more than 70% of the electricity [2]. This coal-dominated energy structure leads to severe problems of greenhouse gas emissions [3,4]. Accordingly, China has proposed many policies to adjust the energy structure in order to manage the challenges related to greenhouse gas emissions. Meanwhile, coal, as a dominant primary energy resource, plays a key role in China's economy. Particularly, coal accounts for about 60% of the total primary energy consumption [2]. Consequently, the fluctuation of coal prices leads to more uncertainty of China's economy through influencing e.g. inflation [5]. Therefore, an accurate forecast for coal prices is important for the policy-makers and governments to make effective strategies which can not only mitigate environmental pollution but also ensure economic

growth. In this paper, we aim to provide a more accurate forecast for steam coal price in China.

Despite the importance of accurate forecast of coal prices, the literature related to coal prices in China is very limited. One possible explanation for the scarcity of related literature is the special regulations on China's coal market in previous years: the double-track pricing system of electrical coals and the key contract whose price is 10%–30% lower than the market price. Compared with coal prices, other energy prices are well studied, which provides us with some references to study coal prices. Recently, many studies document that considering related factors, especially the effective information in high-frequency factors, can significantly improve the forecast accuracy for energy prices [6–8]. However, these studies only focus on the point forecasts, which only provide a single value of the predicted points without any information about the uncertainty of the forecasted values [9]. Most recently, density forecast is attracting more attention in forecasting energy prices, e.g. oil returns [10], because density forecast can provide the uncertainty of the point forecasts by constructing a probability density function of the forecast results [11].

The probability density forecasts for steam coal prices are more useful to policy-makers and power plants because the probability density forecasts are able to provide information about the

* Corresponding author.

E-mail address: m.han@rug.nl (M. Han).

probability or uncertainty of the forecasted coal prices. With such information, policy-makers can formulate more reasonable strategies about how to control the supply, demand, import or export of steam coal and other energies, and how to encourage the utilization of renewable energy sources in power plants. Thus, the government can make a better balance between economic growth and environmental pollution. On the other hand, the information of uncertainty of future coal prices helps power plants to make a better plan to store steam coal, which ensures reliable power generation.

Therefore, this paper provides probability density forecasts for steam coal prices in China considering high-frequency drivers as well. We forecast weekly steam coal prices with daily related factors for three reasons. First, the planning horizon for power plants typically spans weeks [12]. So the forecast for weekly coal prices is more useful to them. Second, weekly prices are able to represent the dynamic characteristics of coal market better than monthly prices which are most widely forecasted in the existing literature because the volatility of monthly prices might be masked by the time aggregation [2,13]. Third, the influence of related drivers on coal prices may only last less than one week, which might not be identified with weekly or monthly data. Hence, we employ daily data for related factors to study their effects on weekly coal prices.

We choose drivers of coal prices from five aspects, including energy, supply, demand, macroeconomics and weather. Particularly, we select Daqing oil, renewable energy source, Japanese natural gas and Australia steam coal prices to proxy energy market, select coal mining industry index as supply factor, select A-share power sector index as demand factor, select A-share index and coal industry index as macroeconomic factors, select temperature to proxy the weather condition. Then, we apply the combined mixed data sampling (C-MIDAS) regression models to study details of the dynamic influence of these daily factors and achieve point forecasts for weekly coal prices because MIDAS model is widely used to match mix-frequency data [7,14,15]. Nevertheless, the C-MIDAS model only reflects the linearity of coal prices. We then combine the extreme gradient boosting (XGBoost) with the C-MIDAS model to construct a hybrid model named C-MIDAS-X model. C-MIDAS-X model allows the XGBoost to represent the nonlinearity of coal prices by correcting the forecast error of the C-MIDAS model. Next, we apply non-parametric bootstrap in the C-MIDAS-X model to obtain bootstrap forecasts for coal prices, which can provide a series of forecasts for each future point in time because it can incorporate parameter uncertainty without any assumption about error distribution [16]. Last, we apply kernel density estimation to the bootstrap forecasts in order to achieve the probability density forecasts for coal prices.

By doing so, we have some interesting results. Different factors have different ways to affect coal prices as expected. Indeed, the influence of temperature lasts longer than other factors. In addition, the Australia steam coal prices, renewable energy source and A-share index are the three best predictors for steam coal prices in China and considering other factors can improve the forecast accuracy for coal prices. Furthermore, the XGBoost performs well in capturing the nonlinearity of coal prices and increases the forecast accuracy of the C-MIDAS model by about 22%. Fourth, our proposed method can provide accurate probability density forecasts for steam coal prices.

Therefore, this paper mainly contributes to the existing literature related to coal prices in three aspects. First, we apply high-

frequency drivers (daily) to study their effects on coal prices at a lower frequency (weekly). So we can identify some influence that might be ignored with a lower frequency of drivers. Second, a novel combination of the extreme gradient boosting (XGBoost) and the C-MIDAS model (C-MIDAS-X) is proposed to forecast coal prices. This method can not only consider the influence of high-frequency drivers but also capture the nonlinearity of coal prices. Third, the full probability density curves of the coal prices are provided by taking advantages of the high-frequency drives. To the best of our knowledge, this paper is the first work to provide the probability density forecast for coal prices.

The remainder of this paper is organized as follows: Section 2 provides a brief literature review about the studies on coal prices. Section 3 introduces the methods that are used in this paper. Section 4 describes the data. Section 5 analyzes the empirical results. Section 6 reports the conclusions of this work.

2. Literature review

In recent years, coal prices have obtained increasing attention. A large strand of the existing literature is about the drivers of coal prices. The coal supply, demand and macroeconomic situation are intuitive factors to affect coal prices. In the Chinese electricity mix, coal-fired power contributes to more than 60% and renewable energy sources contribute to about 26%, while other energies contribute to about 6% [17]. Therefore, other energies, as substitutions of steam coal, are considered as drivers of coal prices. Yang et al. [18] declare a high relationship between coal and oil prices. Hasan and Ratti [19] investigate the influence of oil prices on coal sector stock return and find that a negative relationship between them. Zamani [20] confirms that the shocks of the supply and demand of oil can affect coal prices. Serletis and Xu [21] reject the hypothesis that natural gas and coal prices are not correlated. Fell and Kaffine [22] find that lower natural gas prices can decrease the coal-fired generation.

Besides, international steam coal prices are expected to influence the steam coal prices in China as well. The world steam coal market is geographically divided into the Pacific market (e.g., China, Japan, South Korea, India, Australia, Vietnam, Russia, and Indonesia) and the Atlantic market (e.g., Russia, Poland, European Union countries, South and North African, and Mediterranean countries) [23,24]. Some evidence shows that the world steam coal market is increasingly integrated, but the integration is not complete yet [25–28]. The Australia coal price is most important to set the coal prices on the Pacific markets [29] and even the world coal prices [26].

Additionally, climate change leads to variations in temperatures, which affects energy demand and thus energy prices by affecting e.g. the demand of electricity [30]. Franco and Sanstad [31] mention that an important way through which the climate change influences the electric power system is to increase cooling demand. Coal and electricity prices are closely related to each other. He et al. [32] show that the increase in coal prices leads to an increase in electricity prices. Liu et al. [33] find a long-term relationship between coal and electricity prices even though this relationship is weak. Cui and Wei [17] conclude that steam coal prices are determined by electricity prices. Given the interaction between coal and electricity prices, the temperature is expected to influence coal prices as well. However, the details about the interaction between

temperature and coal prices remain unclear.

Another small strand of the literature is about the forecast for coal prices. Among these limited studies, most of them only consider coal price itself. Zhang and Ma [34] extract the principal components of the coal price index in the past twenty days to forecast the index level in the next day with partial least-squares regressions. Zhao et al. [35] use the multi-fractal detrended fluctuation analysis to study the fluctuations of steam coal prices and propose a quarterly fluctuation index to forecast coal prices. Alameer et al. [29] apply a deep learning model to forecast coal prices at different horizons. Besides, only very few studies consider the influence of related drivers when forecasting coal prices. Guo et al. [36] apply vector autoregression model to forecast coal prices by considering historical coal prices and some related factors such as port stocks, futures prices, sales volume, crude oil price, and producer price index.

However, these studies apply the same frequency for the drivers and coal prices. How the high-frequency drivers influence the low-frequency coal prices remains unclear. Additionally, whether these high-frequency factors can improve the forecast accuracy for coal prices remains unknown. Besides, these studies are all about the point forecasts for coal prices, while the probability density forecasts are more useful to policy-makers because the probability density forecasts can provide uncertainty information about future coal prices. In order to fill these gaps, we select drives of coal prices from five aspects, e.g., energy, supply, demand, macroeconomics and weather. We use daily data for the drivers to give more insights into their influence on weekly coal prices. Then we test whether these high-frequency drivers are useful to forecast coal prices. With the effective information in the high-frequency drivers, we provide the probability density forecasts for coal prices.

3. Methodology

3.1. C-MIDAS-X model regression model

This section demonstrates the C-MIDAS-X model, which is a combination of the combined mixed data sampling (C-MIDAS) regression models and the extreme gradient boosting (XGBoost). Therefore, the C-MIDAS-X model has the advantages of both the C-MIDAS and XGBoost models. Particularly, the C-MIDAS-X model can use high-frequency factors to forecast coal prices at a lower frequency, which exploits the effective information in high-frequency factors [12]. Furthermore, one model cannot always keep better performance [12]. The C-MIDAS-X model combines the best single MIDAS models with individual factors and thus performs well even under structural breaks. Additionally, this model can capture the nonlinearity of coal prices as well, which is well documented in the literature [37]. The details of the C-MIDAS-X model are shown in the following subsections.

3.1.1. C-MIDAS regression model

The mixed data sampling (MIDAS) regression models are proposed firstly to address the mixed-frequency data in financial field [38]. Recently, the MIDAS models have shown good performance in the studies about energy and environmental problems as well [6,39]. The combined mixed data sampling (C-MIDAS) regression models are proposed by combining the single MIDAS models because the forecast combinations can improve the forecast accuracy [7].

The forecast result of the C-MIDAS model based on N single MIDAS models with individual factor is defined as:

$$\hat{y}_{c,t} = \sum_{j=1}^N \hat{w}_j \hat{y}_{j,t} \quad (1)$$

where $\hat{w}_j$ is the combination weight of the forecast results $\hat{y}_{j,t}$. $\hat{y}_{j,t}$ represents the forecast at time t of the best single MIDAS model, which has the highest forecast accuracy with the j^{th} individual factor.

There are several forms for the combination weights. According to Zhao et al. and Han et al. [12,39], we consider five weight functions. (1) Mean squared forecast error (MSFE)-weight: $w_j = m_j^{-1} / \sum_{j=1}^N m_j^{-1}$, where $m_j = \sum_{i=T_0}^t (\delta^{i-T_0} (y_i - \hat{y}_{j,i}))^2 / (t - T_0 + 1)$, $t - T_0 + 1$ is the number of the out-of-sample observations, and $\delta = 1$. (2) Discounted mean squared forecast error (DMSFE)-weight: $\delta = 0.9$ in the MSFE-weight. (3) AIC-weight: $w_j = \exp(-AIC_j) / \sum_{j=1}^N \exp(-AIC_j)$. (4) BIC-weight: $w_j = \exp(-BIC_j) / \sum_{j=1}^N \exp(-BIC_j)$. (5) Equal weight: $w_j = 1/N$.

Next key step to establish the C-MIDAS model is to determine the best single MIDAS model with individual factor. We consider the MIDAS model incorporating the lag effect of individual factor and the autoregressive effect of coal prices, namely $AR - MIDAS(m, K, h)$, which is defined as:

$$y_t = \alpha + \sum_{i=1}^p \gamma_i y_{t-i} + \beta W(L^{1/m}, \theta) x_{t-h/m}^{(m)} + \varepsilon_t \quad (2)$$

where y_t denotes the weekly coal prices, α , γ_i and β are unknown parameters, p is the maximum lag order for the coal prices, and $x_t^{(m)}$ is a daily individual factor that is observed 5 times in a week, that is $m = 5$. $t = 1, 2, \dots, T$, where T is the number of the observations in the estimation sample. Here, h means step ahead. When $h = 0$, we use the daily factor on Friday and before to explain the coal prices for this week. When $h = 1$, we use the daily factor on Thursday and before to explain the coal prices for this week. $W(L^{1/m}, \theta)$ is decided by $L^{1/m}$ and a parameter vector with limited dimension θ . Set $L^{k/m} x_t^{(m)} = x_{t-k/m}^{(m)}$ where $L^{1/m}$ refers to a lag operator. When $k = 0$, $x_{t-k/m}^{(m)}$ refers to the factor on Friday and when $k = 1$, $x_{t-k/m}^{(m)}$ refers to the factor on Thursday. Then, $W(L^{1/m}, \theta)$ can be defined as $W(L^{1/m}, \theta) = \sum_{k=0}^K w(k; \theta) L^{k/m}$, where $w(k; \theta)$ is a polynomial weight, and K is the maximum lag order of individual factor.

Regarding the polynomial weight, we consider the beta polynomial (Beta), which is defined as $w(k; \theta) = w(k; \theta_1, \theta_2) = f(k/K, \theta_1, \theta_2) / \sum_{k=1}^K f(k/K, \theta_1, \theta_2)$ given $\Gamma(\theta) = \int_0^1 e^{-x} x^{\theta-1} dx$ and $f(x_i, \theta_1, \theta_2) = x_i^{\theta_1-1} (1-x_i)^{\theta_2-1} \Gamma(\theta_1 + \theta_2) / \Gamma(\theta_1) \Gamma(\theta_2)$; BetaNN polynomial (BetaNN), which refers to $w(k; \theta) = w(k; \theta_1, \theta_2, \theta_3) = f(k/K, \theta_1, \theta_2) / \sum_{k=1}^K f(k/K, \theta_1, \theta_2) + \theta_3$ given $\theta_1 = 1$ and the same definitions of $f(x_i, \theta_1, \theta_2)$ and $\Gamma(\theta)$ with beta polynomial; Exponential Almon lag polynomial (ExpAlmon), which is $w(k; \theta) = e^{(\theta_1 k + \theta_2 k^2 + \dots + \theta_p k^p)} / \sum_{k=1}^K e^{(\theta_1 k + \theta_2 k^2 + \dots + \theta_p k^p)}$; Almon lag polynomial (Almon), which refers to $\beta w(k; \theta_0, \theta_1, \theta_2, \theta_3) = \sum_{p=0}^3 \theta_p k^p$; Step function polynomial (Step), which is defined as $\beta w(k; \theta) = \theta_1 I_{i \in [a_0, a_1]} + \sum_{p=2}^P \theta_p I_{i \in [a_{p-1}, a_p]}$, $I_{i \in [a_{p-1}, a_p]}$ equals to one when $a_{p-1} \leq i \leq a_p$, otherwise it equals to zero, and $a_0 = 1 < a_1 < \dots < a_p = K$; Unrestricted weight function (UMIDAS), which

generalizes to $y_t = \alpha + B(\beta, L^{1/m})x_t^{(m)} + \varepsilon_t$ with $B(\beta, L^{1/m}) = \sum_{k=0}^K \beta_k L^{k/m}$.

3.1.2. Extreme gradient boosting

The XGBoost model is applied to capture the nonlinearity of coal prices which cannot be captured by the C-MIDAS model by correcting the forecast errors of the C-MIDAS models. XGBoost is an improved boosting algorithm based on the gradient boosting decision tree [40]. Let e_t represents the forecast errors of the C-MIDAS models at time t . The predicted value for the e_t by the XGBoost model is defined as:

$$\hat{e}_t = \sum_{z=1}^Z f_z(g_t) \quad (3)$$

where g_t represents the t^{th} eigenvector. Z refers to the number of trees. $f(\cdot)$ is a regression tree in tree space. These optimal parameters are decided by minimizing the regularized objective function:

$$L(\phi) = \sum_{t=1}^n l(\hat{e}_t, e_t) + \sum_{z=1}^Z \Omega(f_z) \quad (4)$$

where L is the regularized objective function. l is the loss function that measures the difference between the predicted value $\hat{e}_t$ and the target value e_t . In this paper, we use the Mean Square Error (MSE) function as the loss function. n is the sample size. $\Omega(f_z)$ represents the complexity of the z^{th} tree model and $\Omega(f_z) = \gamma T + 0.5\lambda\|w\|^2$. T is the number of the leaves in a tree, γ controls the complexity penalty for tree structure on T , w is the weight of the leaf, and λ controls the degree of the regularization of f_z .

Finally, the final forecast of the C-MIDAS-X model is the sum of the forecast of the C-MIDAS model and the error correction $\hat{e}_t$, which is defined as $\hat{y}_t = \hat{y}_{c,t} + \hat{e}_t$. Plugging in equations (1) and (2), we define the final forecast of the C-MIDAS-X model as:

$$\hat{y}_t = \sum_{j=1}^N \hat{w}_j \left(\hat{\alpha}_j + \sum_{i=1}^p \hat{\gamma}_{ij} y_{t-i} + \hat{\beta}_j W \left(L^{1/m}, \hat{\theta}_j \right) x_{t-h/m,j}^{(m)} \right) + \hat{e}_t \quad (5)$$

3.2. Non-parametric bootstrap and kernel density estimation

3.2.1. Non-parametric bootstrap

Based on the estimations of the C-MIDAS-X model, we conduct a non-parametric bootstrap process because it does not depend on a prior assumption about the distribution of the final forecast error of the C-MIDAS-X model that can be calculated as $\varepsilon_t = y_t - \hat{y}_t$. This non-parametric bootstrap is based on obtaining bootstrap replicates through the following recursion:

$$y_t^* = \sum_{j=1}^N \hat{w}_j \left(\hat{\alpha}_j + \sum_{i=1}^p \hat{\gamma}_{ij} y_{t-i}^* + \hat{\beta}_j W \left(L^{1/m}, \hat{\theta}_j \right) x_{t-h/m,j}^{(m)} \right) + \hat{e}_t + \varepsilon_t^* \quad (6)$$

where ε_t^* is obtained by sampling with replacement from the

empirical distribution function of ε_t . Therefore, we can generate a bootstrap series y_t^* for coal prices. Then the bootstrap forecasts are obtained by estimating equation (5) based on the bootstrap series y_t^* for the coal prices and the original factors. By repeating this procedure 999 times, we can obtain 999 bootstrap point forecasts for each out-of-sample point in time.

3.2.2. Kernel density estimation

Based on the bootstrap forecasts, we apply the kernel density estimation model to simulate the true probability distribution of the forecast results for each out-of-sample point in time. This model is a non-parametric method to estimate probability density by a smooth spike function [41,42]. Denote that $\hat{y}_{t,1}^*, \hat{y}_{t,2}^*, \dots, \hat{y}_{t,n}^*$ are the bootstrap forecasts at time t . n refers to the number of bootstrap repetitions, thus $n = 999$. The probability density function is estimated as:

$$\hat{f}_b(s_t) = \frac{1}{n} \sum_{i=1}^n K_b \left(s_t - \hat{y}_{t,i}^* \right) = \frac{1}{nb} \sum_{i=1}^n K_b \left(\frac{s_t - \hat{y}_{t,i}^*}{b} \right) \quad (7)$$

where b refers to a bandwidth, which is a smoothing parameter and requires self-actuated setting. $K(\cdot)$ is a non-negative kernel function. This paper considers the Gaussian kernel, which is defined as $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$, since it is the most commonly used kernel function due to its good performance [41,42]. The full probability density for the bootstrap forecasts at time t can be generated by repeating the density estimation for a range of different values for s_t .

4. Data

In this paper, we forecast weekly steam coal prices in China with daily factors from five aspects, including energy, supply, demand, macroeconomics and weather. As to the steam coal prices, we use the steam coal Free-on-Board (FOB) prices in Qinhuangdao Port in line with Fan et al. and Zhao et al. [35,37]. This port is the largest coal shipping port in China. So the coal prices in Qinhuangdao Port can reflect the coal prices in China efficiently.

As to energy-side factors, we select renewable energy source, Daqing oil, Japanese natural gas and Australia steam coal prices. We consider renewable energy source because it contributes to 26% of electricity generation and its contribution will increase in future because of less pollution and recyclability [41]. We use renewable energy index price (RES) to represent the renewable energy source. We also consider oil and natural gas even though their contribution in electricity generation is less than steam coal and renewable energy because there do exist some papers suggest the relationship between them and coal prices [18,20,22]. Regarding oil, we select Daqing crude oil spot prices, which is representative because the Daqing oilfield is the largest oilfield in China [43]. For natural gas, we choose Japanese natural gas continuous futures prices for two reasons. On the one hand, the natural gas market in China is not mature yet. On the other hand, the Japanese market is one of the largest international gas markets and is able to affect the global trade market [44]. Additionally, China is a net importer of coal and the import is mainly from Australia [29]. Therefore, we also consider Australia Newcastle steam coal FOB prices (NEWC).

In order to proxy the coal supply, we select coal mining industry

Table 1
Summary statistics of the growth rates of variables.

Variable	Mean	Std.	Min	Max	Obs.
Steam Coal	-0.0236	1.8278	-6.3013	7.5050	373
RES	0.0237	1.7532	-9.8277	6.4758	1865
Daqing Oil	-0.0431	2.1854	-11.5218	13.1344	1865
Japanese Gas	-0.0956	2.1284	-26.7260	35.4382	1865
NEWC Coal	-0.0184	1.2702	-12.7255	10.3661	1865
Supply	-0.0350	1.9529	-10.1556	9.0179	1865
Demand	0.0127	1.4426	-8.1277	6.7557	1865
A-share	0.0126	1.3359	-8.8694	5.5989	1865
Coal Index	-0.0291	1.9318	-9.5175	8.8313	1865
Temp	0.0410	5.3282	-30.6439	45.2039	1865

Notes: The sample starts from January 07, 2013 to February 28, 2020. Steam coal refers to steam coal FOB price in Qinhuangdao Port. RES refers to the renewable energy source. Daqing Oil refers to Daqing crude oil spot prices. Japanese Gas refers to Japanese natural gas continuous futures prices. NEWC Coal refers to Australia Newcastle steam coal FOB prices. Supply refers to the coal mining industry index. Demand refers to the A-share power sector index. A-share refers to the A-share index. Coal Index refers to coal industry index. Temp refers to the average temperature in China.

index to represent the production level of coal. As to the coal demand, we consider A-share power sector index because steam coal is mainly used to generate electricity. As to the macroeconomic situation, we select the A-share index, which is a leading stock index in China. We also consider the coal industry index to reflect the general development of the coal industry. Finally, we select the average temperature in China to proxy the weather condition, in line with Han et al. [39]. The Japanese natural gas futures price is available in Eikon dataset, the temperature is available in the National Oceanic and Atmospheric Administration (NOAA), and other factors are available in Wind dataset.

Therefore, our dataset includes the daily selected factors and weekly steam coal prices. Our sample starts from January 07, 2013 to February 28, 2020. This is because the double-track pricing system of electrical coals and the key contract have been entirely cancelled since 2013, which marks that the coal prices have been market-oriented since then. The estimation sample is from January 07, 2013 to October 04, 2019, which is used to estimate the proposed models. The out-of-sample is from October 11, 2019 to February 28, 2020, whose length is decided according to Ghysels and Ozkan, and Han et al. [39,45]. In order to eliminate the heteroscedasticity, we consider the growth rates or returns of the variables in the empirical analysis. They are calculated as:

$$growth_{it} = \ln(value_{it} / value_{it-1}) \times 100 \quad (8)$$

where i refers to steam coal prices, renewable energy source, Daqing oil, Japanese natural gas, Australia steam coal prices, coal mining industry index, A-share power sector index, A-share index, coal industry index, and the temperature.

Table 1 shows the summary statistics of the growth rates of variables. We find that the Japanese natural gas has the smallest average return, while Australia Newcastle steam coal return is less volatile, followed by A-share index return. The changing trends of these variables are also plotted in Fig. 1.

5. Empirical results

This section analyzes the influence of the high-frequency factors and their usefulness to forecast steam coal prices by comparing root mean squared errors (RMSE) of the proposed models.¹ RMSE is a

¹ $RMSE = \sqrt{\frac{1}{S} \sum_{t=1}^S (y_t - \hat{y}_t)^2}$, where S is the number of out-of-sample data. y_t and $\hat{y}_t$ are the actual and forecasted steam coal prices at time t .

good indicator of out-of-sample performance and smaller RMSE represents better performance [39]. Based on the influence of the high-frequency factors, we provide a probability density forecast for steam coal prices. Additionally, in order to check the robustness of our results to evaluation indicator, we also use mean absolute error (MAE) to evaluate the model performance,² in line with Ahmed et al., He et al., and Zhang et al. [9,46,47]. Smaller MAE represents higher forecast accuracy.

5.1. The influence of the high-frequency factors on coal prices

We apply the single MIDAS models, introduced in equation (2) in section 3.1.1, to separately study the effects of daily factors on weekly steam coal prices. One key step is to decide the best polynomial weight among the six different types discussed in section 3.1.1. Unlike previous studies which preset the polynomial weight before analysis [48], we compare the performance of the six polynomial weights and chooses the one that best captures the character of the relationship between steam coal prices and a single factor, while the predetermined polynomial weight may fail to represent this relationship. We estimate the single MIDAS models with different polynomial weights with the estimation sample and then test their out-of-sample performance according to RMSE.

The optimal maximum lag orders for a single factor and steam coal prices are also determined by comparing RMSE. In order to display the changing trend of RMSEs, we set the maximum lag orders for single factor ranging from 1 to 59 days and the maximum lag orders for steam coal prices ranging from 0 to 5 weeks. Besides, we use the optimal step ahead to reflect the sensitivity of steam coal prices to a single factor. If the factor on Friday needs to be considered to forecast steam coal prices in this week, namely $h = 0$, the steam coal prices are very sensitive to this factor and can react immediately to the shock of this factor. If the factor on Friday does not help to forecast the steam coal prices in this week, while the factor on Thursday performs better, it takes 1 day for steam coal prices to react to the shock of this factor. Particularly, we set the step ahead changing from 0 to 4 days.

In this paper, we take the condition in which we set the value of step ahead as 3 days for Japanese natural gas prices to explain how to determine the best polynomial weight and the optimal maximum lag orders for Japanese natural gas prices and steam coal prices. The results are shown in Table 2. In the condition that the lag order for steam coal prices is 0, the single MIDAS model with the ExpAlmon weight and 8-lag order for the daily natural gas prices performs best. That is the AR(0)-ExpAlmon-MIDAS(5,8,3) has the best performance among all the single MIDAS models without considering the autoregressive effect of steam coal prices because this model has the smallest RMSE under this condition. Similarly, we select the single MIDAS models that have the best performance under the conditions with different lag orders for steam coal prices. These best single MIDAS models, bolded in Table 2, suggest that the ones considering the autoregressive effect of steam coal prices perform better than the one without considering this autoregressive effect. The RMSEs are decreased by at least 26%. This finding provides evidence that steam coal prices are autocorrelated. Moreover, the single MIDAS model with 5-lag order for steam coal prices, 20-lag order for natural gas prices, and ExpAlmon type weight, namely AR(5)-ExpAlmon-MIDAS(5,20,3), is the best one among all the MIDAS models under the condition of $h = 3$ because this model has the lowest RMSE.

With the similar mechanism discussed above, we choose the

² $MAE = \frac{1}{S} \sum_{t=1}^S |y_t - \hat{y}_t|$, where S is the number of out-of-sample data. y_t and $\hat{y}_t$ are the actual and forecasted steam coal prices at time t .

Fig. 1. The changing trends of the growth rates of variables.

Table 2
RMSEs of the single MIDAS models with Japanese natural gas in the condition of $h = 3$.

Polynomial weights	Lag orders of the natural gas						
	8	14	20	30	40	50	59
Panel A: 0-lag order for steam coal prices							
Beta	0.8454	1.0185	1.1473	1.1222	1.1205	1.1300	1.1426
BetaNN	0.8174	0.9383	1.1484	1.1173	1.2329	1.2622	1.1653
ExpAlmon	0.7714	1.0622	0.8938	1.1289	1.1459	1.1508	0.8581
Almon	0.7908	1.0646	1.1456	1.1100	1.2016	1.1429	1.1872
Step	0.8287	1.1007	1.2337	1.2530	1.3990	1.3925	1.4584
UMIDAS	0.8620	0.8434	1.0692	1.3144	1.1910	1.2567	1.3487
Panel B: 1-lag order for steam coal prices							
Beta	0.5954	0.6397	0.6843	0.6790	0.6781	0.6777	0.6775
BetaNN	0.5648	0.5842	0.6319	0.6198	0.6650	0.6779	0.6892
ExpAlmon	0.5633	0.6592	0.5583	0.6799	0.5961	0.6799	0.6462
Almon	0.5932	0.6452	0.7352	0.6623	0.7653	0.6547	0.6773
Step	0.6278	0.7488	0.8535	0.8841	0.9755	1.0140	1.1581
UMIDAS	0.6161	0.6432	0.8188	1.2015	1.3708	1.5102	1.5674
Panel C: 2-lag order for steam coal prices							
Beta	0.5932	0.6376	0.6855	0.6809	0.6802	0.6800	0.6798
BetaNN	0.5631	0.5842	0.6348	0.6229	0.6695	0.7323	0.6949
ExpAlmon	0.5615	0.6606	0.5576	0.6816	0.5940	0.6816	0.6470
Almon	0.5904	0.6463	0.7369	0.6667	0.7665	0.6562	0.6820
Step	0.6260	0.7488	0.8540	0.8817	0.9785	1.0123	1.1616
UMIDAS	0.6145	0.6439	0.8212	1.2048	1.3709	1.5132	1.5644
Panel D: 3-lag order for steam coal prices							
Beta	0.5958	0.6436	0.6721	0.6697	0.6693	0.6690	0.6819
BetaNN	0.5666	0.5902	0.6865	0.6029	0.6685	0.6619	0.6856
ExpAlmon	0.5666	0.6591	0.6712	0.6714	0.5968	0.6715	0.6466
Almon	0.5843	0.6445	0.7168	0.6553	0.7697	0.6418	0.6600
Step	0.6259	0.7375	0.8335	0.8441	0.9548	0.9768	1.1043
UMIDAS	0.6076	0.6374	0.7948	1.1503	1.3931	1.5367	1.5552
Panel E: 4-lag order for steam coal prices							
Beta	0.5918	0.6437	0.6676	0.6641	0.6436	0.6419	0.6412
BetaNN	0.5654	0.5815	0.6141	0.5971	0.6841	0.6981	0.6946
ExpAlmon	0.5660	0.6508	0.5445	0.6741	0.6741	0.6653	0.6480
Almon	0.5780	0.6376	0.6878	0.6456	0.7754	0.6470	0.6732
Step	0.6131	0.7254	0.8139	0.8365	0.9895	1.0051	1.1424
UMIDAS	0.5958	0.6333	0.7710	1.1284	1.3818	1.4862	1.4746
Panel F: 5-lag order for steam coal prices							
Beta	0.5919	0.6437	0.6666	0.6632	0.6447	0.6426	0.6418
BetaNN	0.5656	0.5810	0.6125	0.5978	0.6870	0.7004	0.6961
ExpAlmon	0.5662	0.6501	0.5436	0.6729	0.6648	0.6403	0.6483
Almon	0.5781	0.6360	0.6869	0.6456	0.7795	0.6484	0.6740
Step	0.6139	0.7247	0.8129	0.8377	0.9898	1.0051	1.1437
UMIDAS	0.5958	0.6345	0.7761	1.1289	1.3837	1.4868	1.4690

Notes: This table shows the RMSEs of the single MIDAS models with Japanese natural gas in the condition of $h = 3$ based on the estimation with the observations in the estimation sample from January 07, 2013 to October 04, 2019 using equation (2) in section 3.1.1. Due to the availability of the space, this table only displays the single MIDAS models with the maximum lag orders for the natural gas as 8, 14, 20, 30, 40, 50, 59. The bold values are the smallest RMSEs of the single MIDAS models under the condition of different maximum lag order for steam coal prices. Beta, BetaNN, ExpAlmon, Almon, Step and UMIDAS represents different polynomial weights discussed in section 3.1.1.

best single MIDAS models with renewable energy source, Daqing oil, Japanese natural gas, Australia steam coal prices, supply, demand, A-share index, coal industry index and the temperature under conditions with a different step ahead, which are bolded Table 3. The results show that the polynomial weights for the best single MIDAS models are different with different individual factors. The Almon weight performs well in capturing the influence of Australia steam coal prices, A-share index and coal industry index. The ExpAlmon weight performs well for Japanese natural gas and temperature. The Step weight performs well for Daqing oil and supply, while the UMIDAS weight performs well in demonstrating the effects of renewable energy source and demand on steam coal

prices. This finding provides evidence that we should consider all the polynomial weights and choose the best one for each factor instead of predetermining the polynomial weight, consistent with the study of Zhao et al. [12].

Moreover, Table 3 shows that different factors have different effects on steam coal prices as expected. In particular, the influence of temperature lasts for 54 days, which is longer than other factors. This might be because temperature significantly affects the residential electricity consumption, especially extreme temperature [42]. For instance, if temperature increase, residents are likely to use more home appliances and thus more electricity. Moreover, more than 70% of the electricity are generated by steam coal [2].

Table 3
The best single MIDAS models with different factors.

H	Model	RMSE	Model	RMSE
Panel A: RES				
0	AR(5)-UMIDAS-MIDAS(5,27,0)	0.5624	Panel B: Daqing Oil	
1	AR(5)-UMIDAS-MIDAS(5,26,1)	0.5460	AR(5)-Step-MIDAS(5,52,0)	0.5540
2	AR(5)-UMIDAS-MIDAS(5,28,2)	0.5399	AR(2)-Step-MIDAS(5,55,1)	0.5595
3	AR(5)-UMIDAS-MIDAS(5,27,3)	0.5542	AR(4)-ExpAlmon-MIDAS(5,60,2)	0.5710
4	AR(5)-UMIDAS-MIDAS(5,27,4)	0.5584	AR(5)-Step-MIDAS(5,49,3)	0.5555
Panel C: Japanese Gas				
0	AR(4)-ExpAlmon-MIDAS(5,23,0)	0.5445	AR(2)-Almon-MIDAS(5,48,4)	0.5559
1	AR(2)-BetaNN-MIDAS(5,10,1)	0.5666	Panel D: NEWC Coal	
2	AR(2)-ExpAlmon-MIDAS(5,9,2)	0.5615	AR(2)-ExpAlmon-MIDAS(5,4,0)	0.4982
3	AR(5)-ExpAlmon-MIDAS(5,20,3)	0.5436	AR(1)-Almon-MIDAS(5,3,1)	0.4889
4	AR(5)-ExpAlmon-MIDAS(5,20,4)	0.5602	AR(2)-Beta-MIDAS(5,3,2)	0.5294
Panel E: Supply				
0	AR(2)-Step-MIDAS(5,4,0)	0.5715	AR(1)-Almon-MIDAS(5,3,3)	0.5771
1	AR(2)-Beta-MIDAS(5,4,1)	0.5720	AR(4)-BetaNN-MIDAS(5,22,4)	0.6726
2	AR(2)-Beta-MIDAS(5,3,2)	0.5716	Panel F: Demand	
3	AR(4)-Almon-MIDAS(5,3,3)	0.5738	AR(2)-UMIDAS-MIDAS(5,20,0)	0.5578
4	AR(4)-BetaNN-MIDAS(5,57,4)	0.6003	AR(2)-UMIDAS-MIDAS(5,19,1)	0.5540
Panel G: A-share				
0	AR(2)-UMIDAS-MIDAS(5,25,0)	0.5491	AR(2)-UMIDAS-MIDAS(5,17,2)	0.5508
1	AR(2)-UMIDAS-MIDAS(5,24,1)	0.5419	AR(2)-UMIDAS-MIDAS(5,16,3)	0.5527
2	AR(5)-Almon-MIDAS(5,3,2)	0.5270	AR(2)-UMIDA-MIDAS(5,21,4)	0.5566
3	AR(2)-Almon-MIDAS(5,3,3)	0.5564	Panel H: Coal Index	
4	AR(2)-UMIDAS-MIDAS(5,21,4)	0.5751	AR(2)-Step-MIDAS(5,4,0)	0.5739
Panel K: Temp				
0	AR(2)-ExpAlmon-MIDAS(5,54,0)	0.5692	AR(1)-Almon-MIDAS(5,3,1)	0.5723
1	AR(4)-BetaNN-MIDAS(5,53,1)	0.5803	AR(2)-Beta-MIDAS(5,3,2)	0.5734
2	AR(2)-BetaNN-MIDAS(5,30,2)	0.5865	AR(1)-Almon-MIDAS(5,3,3)	0.5761
3	AR(1)-ExpAlmon-MIDAS(5,29,3)	0.5823	AR(2)-BetaNN-MIDAS(5,60,4)	0.5989
4	AR(2)-BetaNN-MIDAS(5,60,4)	0.5878		

Notes: This table displays the best single MIDAS models with individual factors under the conditions with different h changing from 0 to 4, where h refers to the step ahead of the respective factor. RES refers to the renewable energy source. Daqing Oil refers to Daqing crude oil spot prices. Japanese Gas refers to Japanese natural gas continuous futures prices. NEWC Coal refers to Australia Newcastle steam coal FOB prices. Supply refers to the coal mining industry index. Demand refers to the A-share power sector index. A-share refers to the A-share index. Coal Index refers to coal industry index. Temp refers to the average temperature in China. Take the AR(4)-ExpAlmon-MIDAS(5,23,0) for example, it refers to the single MIDAS model with 4-lag order for steam coal prices, 23-lag order for the factor, ExpAlmon polynomial weight, and $h = 0$. The other model specifications have a similar meaning. The bolded value represents the smallest RMSE under different conditions.

Table 4
RMSEs of the C-MIDAS and C-MIDAS-X models with five combination weights.

Weight	C-MIDAS	C-MIDAS-X
MSFE	0.5234	0.4081
DMSFE	0.5234	0.4560
AIC	0.5648	0.4465
BIC	0.5713	0.4459
Equal weight	0.5262	0.4422

Notes: This table displays the RMSEs of the C-MIDAS and C-MIDAS-X models with five combination weights. The bolded value represents the smallest RMSE among those with different combination weights.

Therefore, it is not surprising that the influence of temperature on steam coal prices is longer than other factors. Besides, the influence of energy-side factors lasts more than 20 days, except Australia steam coal prices, because these energy-side factors are closely related to steam coal prices as substitutes for coal to generate electricity. Compared with supply, demand has longer influence, lasting for 17 days, while the effect of supply only lasts for 4 days.

The influence of the A-share index and coal industry index only lasts for 3 days. The influence less than 1 week might not be identified with data at weekly frequency, which suggests the use of daily data in forecasting weekly steam coal prices. Generally, these findings have important implications for policy-makers. Different factors have different influence on steam coal prices. Some effects last for more than 50 days, while some effects last for less than 1 week. Therefore, policy-makers should make a reasonable plan about when to pay attention to these factors in order to ensure the stability of steam coal prices.

Additionally, Table 3 also shows that all the best single MIDAS models with different factors have autoregressive effect of steam coal prices. Therefore, there exists auto-correlation among steam coal prices, which supports those studies that forecast coal prices only considering the coal prices itself [29,34,35]. Furthermore, the AR(1)-Almon-MIDAS(5,3,1) model with Australia steam coal prices keeps the best performance. This finding suggests that the best predictor of steam coal prices is Australia steam coal prices, which makes sense because China is the largest coal importing country and the import is mainly from Australia [29]. A-share index and renewable energy source also perform comparably to Australia

Fig. 2. The probability density forecasts for steam coal prices in out-of-sample.

Fig. 2. (continued).

steam coal prices in forecasting steam coal prices in China. Additionally, steam coal prices are more sensitive to Daqing oil, supply and temperature, followed by Australia steam coal prices, coal industry index, renewable energy source, demand, A-share index and Japanese natural gas. This sensitivity is reflected by the step ahead in single MIDAS models.

5.2. The usefulness of the high-frequency factors in forecasting coal prices

After analyzing the influence of high-frequency factors separately, this section studies the usefulness of high-frequency factors in forecasting steam coal prices. According to the analysis in section 5.1, we determine the best single MIDAS models with individual factors that are shown in Table 3. Then we combine those best single MIDAS models to forecast steam coal prices with the C-MIDAS model introduced in equation (1) in section 3.1.1. Therefore, we can study whether considering all the selected factors improve the forecast accuracy for steam coal prices by comparing RMSE of the C-MIDAS models with that of the best single MIDAS model with individual factors. Furthermore, we propose a C-MIDAS-X model which combines C-MIDAS and XGBoost models to forecast steam coal prices. The C-MIDAS-X model can reveal the nonlinearity of steam coal prices that are well documented in the existing literature [37].

The forecast results of the C-MIDAS model and the C-MIDAS-X model are shown in Table 4. The C-MIDAS model with MSFE and DMSFE perform better than the ones with other combination weights. Compared with the best single MIDAS models with individual factors shown in Table 3, the C-MIDAS models with MSFE and DMSFE perform better than the best single MIDAS models except for the one with Australia steam coal prices. These results suggest that considering all the high-frequency factors can improve

the forecast accuracy for steam coal prices. This is because the C-MIDAS model can take advantage of the effective information in high-frequency factors captured by the best single MIDAS models [7,49]. Therefore, the high-frequency factors are useful to forecast steam coal prices. These findings provide evidence that we should consider relevant drivers when forecasting steam coal prices instead of only considering coal prices itself, e.g. the research of Guo et al. [36]. Yet, the related studies are very limited.

Even though the C-MIDAS models can improve the forecast accuracy for steam coal prices by considering the high-frequency factors, they can only reveal the linearity of steam coal prices. In order to capture their nonlinearity, we propose the C-MIDAS-X model by applying the XGBoost model to correct the forecast error of the C-MIDAS models, which is shown in section 3.1.2. The forecast results of the C-MIDAS-X shown in Table 4 suggest that the XGBoost can significantly improve the forecast accuracy of the C-MIDAS model in all cases of the five combination weights. Especially, the C-MIDAS-X model with the MSFE combination weight performs best among the five combination weights and improve the forecast accuracy by around 22% compared with the C-MIDAS model with the MSFE combination weight. Therefore, the C-MIDAS-X model can not only consider the influence of high-frequency factors but also capture the nonlinearity of steam coal prices.

5.3. The probability density forecasts for coal prices

Even though the C-MIDAS-X model can improve the forecast accuracy for steam coal prices, it can only provide the point forecast, while the probability density forecast is more important because it can capture the prediction uncertainty [42]. In this section, we provide the probability density forecasts for steam coal prices based on the C-MIDAS-X with MSFE combination weight. As introduced in section 3.2, we apply non-parametric bootstrap with

equation (6) in the C-MIDAS-X model. By replicating the bootstrap process 999 times, we obtain 999 forecasts for each out-of-sample point in time. We further derive the probability distributions of the bootstrap forecasts with kernel density estimation, as shown in equation (7).

Fig. 2 shows the probability density forecasts for steam coal prices in out-of-sample from October 11, 2019 to February 28, 2020. The actual coal prices are also displayed in the probability density curves, which are represented by the vertical lines. We find that the true values of steam coal prices are positioned around the highest probability point in the probability density curves in most cases. Only the true values for the weeks of November 15, 2019, January 10, 2020 and February 28, 2020 appear further from the highest probability point. Therefore, we suggest that our proposed method can provide accurate probability density forecasts for steam coal prices and is effective in quantifying the uncertainty of steam coal prices. The satisfying performance of the probability density forecast is partly due to the good performance of the C-MIDAS-X model which provides point forecasts for steam coal prices.

With these probability density forecasts, policy-makers can make reasonable plans according to the uncertainty of forecasted steam coal prices. Such uncertainty information can help policy-makers, coal enterprises and power plants have a deeper understanding of steam coal prices. Therefore, policy-makers can effectively control the coal supply and demand by making reasonable arrangements about coal production, export or import. Power plants can also reasonably adjust their purchase plan of steam coal. Moreover, the density forecasts of steam coal prices are important to policy-makers to encourage the use of renewable energy in power generation. Therefore, policy-makers can make a balance between economic development and environmental pollution.

5.4. The robustness check

In order to check the robustness of the results discussed above, we provide a robustness check with an alternative indicator to evaluate the performance of the proposed models. In addition to RMSE, we also use mean absolute error (MAE) to evaluate the model performance, in line with Ahmed et al., He et al., and Zhang et al. [9,46,47]. The results are shown in Appendix A and B, which suggests that the results are robust to the selection of evaluation indicator.

The best single MIDAS models with individual factors selected with MAE indicator are shown in Table A1 in Appendix A. Still, the influence of temperature on steam coal prices lasts longer than other factors. The influence of energy-side factors lasts more than 20 days, except for Australia steam coal prices. The steam coal prices have auto-correlation reflected by the best single MIDAS models. Renewable energy source, Australia steam coal prices and A-share index are the three best predictors of steam coal prices because the MAEs of the three best single MIDAS models of these three factors are smaller than other factors. Furthermore, steam coal prices are more sensitive to Daqing oil, Australia steam coal prices and temperature, followed by A-share index, renewable energy source, demand, Japanese natural gas, supply and coal industry index. These results are quite robust.

As shown in Table A2 in Appendix A, the C-MIDAS model with MSFE weight performs better than almost all single MIDAS models with individual factors, except renewable energy source. The C-MIDAS-X model with the DMSFE combination weight perform best and improve the forecast accuracy by around 10% compared with

the C-MIDAS model with the DMSFE combination weight. Therefore, the high-frequency factors can improve the forecast accuracy and the XGBoost model can capture the nonlinearity of steam coal prices.

Based on the forecast results of C-MIDAS-X model selected with MAE, we provide the probability density forecasts for steam coal prices, as shown in Fig. B1 in Appendix B. The actual steam coal prices are positioned around the highest probability point in the probability density curves in most cases, except for November 08, 2019, November 15, 2019, December 27, 2019, January 10, 2020, January 31, 2020 and February 28, 2020. Therefore, the probability density forecasts based on the MIDAS models selected with MAE are also satisfying.

6. Conclusions

In this paper, we aim to provide probability density forecasts for steam coal prices in China based on high-frequency factors. In order to achieve this goal, we apply a mixed-frequency dataset that includes weekly steam coal prices and daily selected factors from energy, supply, demand, macroeconomics and weather, such as renewable energy source, Daqing oil, Japanese natural gas, Australia steam coal prices, coal mining industry index, A-share power sector index, A-share index, coal industry index, and temperature. We separately analyze the influence of daily factors on weekly steam coal prices with single MIDAS models. Then we combine these single MIDAS models (C-MIDAS) to study the usefulness of these daily factors in forecasting weekly steam coal prices. However, the C-MIDAS model can only reveal the linearity of steam coal prices. In order to capture the nonlinearity, we develop C-MIDAS-X model that combines the C-MIDAS and the XGBoost models to forecast steam coal prices by applying the XGBoost model to correct the forecast error of C-MIDAS model. Based on the forecast performance of the C-MIDAS-X model, we apply non-parametric bootstrap process and kernel density estimation model to provide probability density forecasts for steam coal prices.

In general, we draw the following conclusions from the empirical results and these results are robust to evaluation indicator. (1) The influence of temperature on steam coal prices lasts longer than other factors. This might be because temperature significantly affects the residential electricity consumption. (2) Renewable energy source, Australia steam coal prices and A-share index are the three best predictors of steam coal prices. (3) Steam coal prices are more sensitive to Daqing oil, supply and temperature, followed by Australia steam coal prices, coal industry index, renewable energy source, demand, A-share index and Japanese natural gas. (4) There exists auto-correlation among the steam coal prices. (5) It can improve the forecast accuracy for steam coal prices to consider all the selected daily factors. (6) The XGBoost performs well in capturing the nonlinearity of coal prices and improve the forecast accuracy of the C-MIDAS model by about 22%. (7) Our proposed method can provide accurate probability density forecasts for steam coal prices.

In future research, we aim to construct a more comprehensive factor system by considering more drivers of steam coal prices, e.g. the air quality. Besides, we will study other methods to better capture the nonlinearity of steam coal prices. Moreover, we will investigate other probability density forecast method to provide more satisfying density forecasts.

Author contribution

Lili Ding: Conceptualization, Supervision, Writing – review & editing. Zhongchao Zhao: Methodology, Software, Writing – original draft. Meng Han: Methodology, Writing – original draft, Writing – review & editing.

Appendix A. Tables

Table A1

The best single MIDAS models with different factors based on MAE.

H	Model	RMSE	Model	RMSE
	Panel A: RES		Panel B: Daqing Oil	
0	AR(2)-UMIDAS-MIDAS(5,270)	0.3974	AR(1)-Step-MIDAS(5,52,0)	0.4341
1	AR(2)-UMIDAS-MIDAS(5,26,1)	0.3909	AR(1)-Step-MIDAS(5,27,0)	0.4078
2	AR(2)-UMIDAS-MIDAS(5,25,2)	0.3822	AR(2)-ExpAlmon-MIDAS(5,60,0)	0.4291
3	AR(1)-UMIDAS-MIDAS(5,24,3)	0.3935	AR(2)-ExpAlmon-MIDAS(5,59,0)	0.4361
4	AR(2)-UMIDAS-MIDAS(5,26,4)	0.3948	AR(2)-Step-MIDAS(5,24,0)	0.4158
	Panel C: Japanese Gas		Panel D: NEWC Coal	
0	AR(2)-ExpAlmon-MIDAS(5,24,0)	0.4330	AR(2)-BetaNN-MIDAS(5,4,0)	0.3847
1	AR(1)-BetaNN-MIDAS(5,10,1)	0.4402	AR(1)-Almon-MIDAS(5,3,1)	0.3852
2	AR(1)-ExpAlmon-MIDAS(5,22,2)	0.4377	AR(2)-Beta-MIDAS(5,3,2)	0.3980
3	AR(2)-ExpAlmon-MIDAS(5,21,3)	0.4324	AR(2)-Beta-MIDAS(5,23,3)	0.4425
4	AR(4)-BetaNN-MIDAS(5,20,4)	0.4141	AR(2)-BetaNN-MIDAS(5,22,4)	0.5083
	Panel E: Supply		Panel F: Demand	
0	AR(1)-Step-MIDAS(5,4,0)	0.4421	AR(2)-UMIDAS-MIDAS(5,25,0)	0.4290
1	AR(1)-Beta-MIDAS(5,4,1)	0.4420	AR(2)-UMIDAS-MIDAS(5,24,1)	0.4233
2	AR(1)-Beta-MIDAS(5,3,2)	0.4422	AR(2)-UMIDAS-MIDAS(5,23,2)	0.4219
3	AR(1)-BetaNN-MIDAS(5,27,3)	0.4413	AR(2)-UMIDAS-MIDAS(5,22,3)	0.4235
4	AR(2)-ExpAlmon-MIDAS(5,21,4)	0.4598	AR(2)-UMIDAS-MIDAS(5,21,4)	0.4258
	Panel G: A-share		Panel H: Coal Index	
0	AR(1)-UMIDAS-MIDAS(5,29,0)	0.4007	AR(1)-Step-MIDAS(5,4,0)	0.4440
1	AR(2)-UMIDAS-MIDAS(5,28,1)	0.3975	AR(1)-Almon-MIDAS(5,3,1)	0.4423
2	AR(2)-UMIDAS-MIDAS(5,27,2)	0.3994	AR(1)-Beta-MIDAS(5,3,2)	0.4432
3	AR(2)-UMIDAS-MIDAS(5,22,3)	0.4203	AR(1)-Almon-MIDAS(5,3,3)	0.4421
4	AR(1)-UMIDAS-MIDAS(5,24,4)	0.4354	AR(2)-Almon-MIDAS(5,21,4)	0.4566
	Panel K: Temp			
0	AR(2)-ExpAlmon-MIDAS(5,31,0)	0.4360		
1	AR(1)-BetaNN-MIDAS(5,53,1)	0.4442		
2	AR(1)-BetaNN-MIDAS(5,31,2)	0.4490		
3	AR(1)-BetaNN-MIDAS(5,30,3)	0.4514		
4	AR(1)-BetaNN-MIDAS(5,29,4)	0.4524		

Notes: This table displays the best single MIDAS models with individual factors under the conditions with different h changing from 0 to 4, where h refers to the step ahead of the respective factor. RES refers to the renewable energy source. Daqing Oil refers to Daqing crude oil spot prices. Japanese Gas refers to Japanese natural gas continuous futures prices. NEWC Coal refers to Australia Newcastle steam coal FOB prices. Supply refers to the coal mining industry index. Demand refers to the A-share power sector index. A-share refers to the A-share index. Coal Index refers to coal industry index. Temp refers to the average temperature in China. Take the AR(4)-ExpAlmon-MIDAS(5,23,0) for example, it refers to the single MIDAS model with 4-lag order for steam coal prices, 23-lag order for the factor, ExpAlmon polynomial weight, and $h = 0$. The other model specifications have a similar meaning. The bolded value represents the smallest MAE under different conditions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by China Scholarship Council (CSC); National Science Foundation of China (71973132); National Social Science Fund of China (19VHQ002); Taishan Scholar Program (tsqn20161014, ts201712014).

Table A2

MAEs of the C-MIDAS and C-MIDAS-X models with five combination weights.

Weight	C-MIDAS	C-MIDAS-X
MSFE	0.4006	0.3812
DMSFE	0.4012	0.3604
AIC	0.4364	0.4004
BIC	0.4367	0.4063
Equal weight	0.4020	0.3950

Notes: This table displays the MAEs of the C-MIDAS and C-MIDAS-X models with five combination weights. The bolded value represents the smallest MAE among those with different combination weights.

Appendix B. Figures

Fig. B1. The probability density forecasts for steam coal prices in out-of-sample based on the MIDAS models selected with MAE.

Fig. B1. (continued).

References

- [1] Wei C, Löschel A, Liu B. An empirical analysis of the CO₂ shadow price in Chinese thermal power enterprises. *Energy Econ* 2013;40:22–31.
- [2] Li J, Xie C, Long H. The roles of inter-fuel substitution and inter-market contagion in driving energy prices: evidences from China's coal market. *Energy Econ* 2019;84:104525.
- [3] Lin B, Ouyang X. Energy demand in China: comparison of characteristics between the US and China in rapid urbanization stage. *Energy Convers Manag* 2014;79:128–39.
- [4] Yu S, Zhu K, Zhang X. Energy demand projection of China using a path-coefficient analysis and PSO–GA approach. *Energy Convers Manag* 2012;53(1):142–53.
- [5] Guo J, Zheng X, Chen Z-M. How does coal price drive up inflation? Reexamining the relationship between coal price and general price level in China. *Energy Econ* 2016;57:265–76.
- [6] Degiannakis S, Filis G. Forecasting oil prices: high-frequency financial data are indeed useful. *Energy Econ* 2018;76:388–402.
- [7] Zhang Y-J, Wang J-L. Do high-frequency stock market data help forecast crude oil prices? Evidence from the MIDAS models. *Energy Econ* 2019;78:192–201.
- [8] Pan Z, Wang Y, Wu C, Yin L. Oil price volatility and macroeconomic fundamentals: a regime switching GARCH-MIDAS model. *J Empir Finance* 2017;43:130–42.
- [9] He Y, Liu R, Li H, Wang S, Lu X. Short-term power load probability density forecasting method using kernel-based support vector quantile regression and Copula theory. *Appl Energy* 2017;185:254–66.
- [10] Plakandaras V, Gupta R, Wong W-K. Point and density forecasts of oil returns: the role of geopolitical risks. *Resour Pol* 2019;62:580–7.
- [11] Yang Y, Li S, Li W, Qu M. Power load probability density forecasting using Gaussian process quantile regression. *Appl Energy* 2018;213:499–509.
- [12] Zhao X, Han M, Ding L, Kang W. Usefulness of economic and energy data at different frequencies for carbon price forecasting in the EU ETS. *Appl Energy* 2018;216:132–41.
- [13] Arouri MEH, Jouini J, Nguyen DK. On the impacts of oil price fluctuations on European equity markets: volatility spillover and hedging effectiveness. *Energy Econ* 2012;34(2):611–7.
- [14] Ding L, Lv Z, Han M, Zhao X, Wang W. Forecasting China's wastewater discharge using dynamic factors and mixed-frequency data. *Environ Pollut* 2019;255:113148.
- [15] Baumeister C, Guérin P, Kilian L. Do high-frequency financial data help forecast oil prices? The MIDAS touch at work. *Int J Forecast* 2015;31(2):238–52.
- [16] Trucíos C, Hotta LK, Ruiz E. Robust bootstrap forecast densities for GARCH returns and volatilities. *J Stat Comput Simulat* 2017;87(16):3152–74.
- [17] Cui H, Wei P. Analysis of thermal coal pricing and the coal price distortion in China from the perspective of market forces. *Energy Pol* 2017;106:148–54.
- [18] Yang C-J, Xuan X, Jackson RB. China's coal price disturbances: observations, explanations, and implications for global energy economies. *Energy Pol* 2012;51:720–7.
- [19] Hasan MZ, Ratti RA. Coal sector returns and oil prices: developed and emerging countries. *Int J Energy Econ Pol* 2015;5(2):515–24.
- [20] Zamani N. The relationship between crude oil and coal markets: a new approach. *Int J Energy Econ Pol* 2016;6(4):801–5.
- [21] Serletis A, Xu L. Volatility and a century of energy markets dynamics. *Energy Econ* 2016;55:1–9.
- [22] Fell H, Kaffine DT. The fall of coal: joint impacts of fuel prices and renewables on generation and emissions. *Am Econ J Econ Pol* 2018;10(2):90–116.
- [23] Papież M, Śmiech S. Dynamic steam coal market integration: evidence from rolling cointegration analysis. *Energy Econ* 2015;51:510–20.
- [24] Papież M, Śmiech S. Causality-in-mean and causality-in-variance within the international steam coal market. *Energy Econ* 2013;36:594–604.
- [25] Warell L. Market integration in the international coal industry: a cointegration approach. *Energy J* 2006;27(1).
- [26] Batten JA, Brzeszczynski J, Ciner C, Lau MC, Lucey B, Yarovaya L. Price and volatility spillovers across the international steam coal market. *Energy Econ* 2019;77:119–38.
- [27] Zaklan A, Cullmann A, Neumann A, Von Hirschhausen C. The globalization of steam coal markets and the role of logistics: an empirical analysis. *Energy Econ* 2012;34(1):105–16.
- [28] Li R, Joyeux R, Ripple RD. International steam coal market integration. *Energy J* 2010;31(3).
- [29] Alameer Z, Fathalla A, Li K, Ye H, Jianhua Z. Multistep-ahead forecasting of coal prices using a hybrid deep learning model. *Resour Pol* 2020;65:101588.
- [30] Ahmed T, Muttaqi K, Agalgaonkar A. Climate change impacts on electricity demand in the state of new South Wales, Australia. *Appl Energy* 2012;98:376–83.
- [31] Franco G, Sanstad AH. Climate change and electricity demand in California. *Climatic Change* 2008;87(1):139–51.
- [32] He Y, Zhang S, Yang L, Wang Y, Wang J. Economic analysis of coal price–electricity price adjustment in China based on the CGE model. *Energy Pol* 2010;38(11):6629–37.
- [33] Liu M-H, Margaritis D, Zhang Y. Market-driven coal prices and state-administered electricity prices in China. *Energy Econ* 2013;40:167–75.
- [34] Zhang B, Ma J. Coal price index forecast by a new partial least-squares regression. *Procedia Engineering* 2011;15:5025–9.
- [35] Zhao Z-y, Zhu J, Xia B. Multi-fractal fluctuation features of thermal power coal price in China. *Energy* 2016;117:10–8.
- [36] Guo X, Shi J, Ren D. Coal price forecasting and structural analysis in China. *Discrete Dynam Nat Soc* 2016;2016.
- [37] Fan X, Wang L, Li S. Predicting chaotic coal prices using a multi-layer perceptron network model. *Resour Pol* 2016;50:86–92.
- [38] Chen X, Ghysels E. News—good or bad—and its impact on volatility predictions over multiple horizons. *Rev Financ Stud* 2011;24(1):46–81.

- [39] Han M, Ding L, Zhao X, Kang W. Forecasting carbon prices in the Shenzhen market, China: the role of mixed-frequency factors. *Energy* 2019;171:69–76.
- [40] Lu H, Ma X, Huang K, Azimi M. Carbon trading volume and price forecasting in China using multiple machine learning models. *J Clean Prod* 2020;249:119386.
- [41] He Y, Qin Y, Wang S, Wang X, Wang C. Electricity consumption probability density forecasting method based on LASSO-Quantile Regression Neural Network. *Appl Energy* 2019;233:565–75.
- [42] Guo Z, Zhou K, Zhang X, Yang S. A deep learning model for short-term power load and probability density forecasting. *Energy* 2018;160:1186–200.
- [43] Zeng S, Nan X, Liu C, Chen J. The response of the Beijing carbon emissions allowance price (BJC) to macroeconomic and energy price indices. *Energy Pol* 2017;106:111–21.
- [44] Wakamatsu H, Aruga K. The impact of the shale gas revolution on the US and Japanese natural gas markets. *Energy Pol* 2013;62:1002–9.
- [45] Ghysels E, Ozkan N. Real-time forecasting of the US federal government budget: a simple mixed frequency data regression approach. *Int J Forecast* 2015;31(4):1009–20.
- [46] Zhang H, Nguyen H, Bui X-N, Nguyen-Thoi T, Bui T-T, Nguyen N, et al. Developing a novel artificial intelligence model to estimate the capital cost of mining projects using deep neural network-based ant colony optimization algorithm. *Resour Pol* 2020;66:101604.
- [47] Ahmed R, Sreeram V, Mishra Y, Arif M. A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization. *Renew Sustain Energy Rev* 2020;124:109792.
- [48] Pan Z, Wang Q, Wang Y, Yang L. Forecasting US real GDP using oil prices: a time-varying parameter MIDAS model. *Energy Econ* 2018;72:177–87.
- [49] Timmermann A. Forecast combinations. *Handb Econ Forecast* 2006;1:135–96.