A hybrid model of modal decomposition and gated recurrent units for short-term load forecasting

Chun-Hua Wang[1], Wei-Qin Li[2]

[1] School of Electronic Engineering, Xi'an Aeronautical Institute, Xi'an, Shaanxi, China

[2] School of Automation and Information Engineering, Xi'an University of Technology, Xi'an Shaanxi, China

Corresponding Author:

Wei-Qin Li[1]

Jinhua South Road No.5, Xi'an, Shaanxi, 710048, China

Email address: wqlee@xaut.edu.cn

Abstract

Electrical load forecasting is important to ensuring power systems are operated both economically and safely. However, accurately forecasting load is difficult because of variability and frequency aliasing. To eliminate frequency aliasing, some methods set parameters that depend on experiences. The present study proposes an adaptive hybrid model of modal decomposition and gated recurrent units (GRU) to reduce frequency aliasing and series randomness. This model uses average sample entropy and mutual correlation to jointly determine the modal number in the decomposition. Random adjustment parameters were introduced to the Adam algorithm to improve training speed. To assess the applicability and accuracy of the proposed hybrid model, it was compared with some state of the art forecasting methods. The results, which were validated by actual data sets from Shaanxi province, China, show that the proposed model had a higher accuracy and better reliability compared to the other forecasting methods.

Introduction

Electrical load forecasting plays an important role in power system dispatching and security detection. However, it is very difficult to accurately forecast load because the electrical load can be affected by the weather and some accidental factors (Mideksa & Kallbekken 2010; Wang et al. 2012). It is, therefore, important to develop an effective load forecasting method that is both reliable and accurate.

Electrical load measurement data can be contaminated by random noise that reduces accurate forecasting performance (Xiao et al. 2007; Li 2020; Ren & Li 2023). Signal processing technologies have been developed to reduce the random noise created during measurement, such as faults in the sensors or power supply equipment failures (Guan et al. 2021; Wang, Yao & Papaethymiou 2023). Filtering

32 methods, such as the wavelet analysis method and the Kalman filtering method, are currently the most

33 common ways of dealing with the random noise in the data (Quilty & Aadamowski 2018; Nobrega &

34 Oliveira 2019). Electrical load is also affected by people's consumption habits and varies drastically

35 between different periods of time, so frequency aliasing, which occurs when the load sequence is not

36 sampled at a high enough rate, is a significant problem and makes it harder to accurately forecast load

37 from the data. Some methods, including empirical mode decomposition (EMD) and variational mode

38 decomposition (VMD), reduce aliasing by decomposing the load series (Li & Chang 2018; Mounir,

39 Ouadi & Jrhilifa 2023; Rayi et al. 2022). The disadvantage of these methods is that they use parameters

40 that depend on experiences.

41 Short-term load forecasting techniques include statistical models (Ren & Li 2023; Lee & Ko 2021;

42 Jin et al. 2021), the machine learning method (Tarmanini et al. 2023; Xie et al. 2020) and the deep

43 learning method. In these methods, the long and short-term memory (LSTM) network can find the

44 evolution characteristics of a time series based on a large number of training samples, resulting in a higher

45 accuracy than traditional machine learning methods (Mokarram et al. 2023; Rafi et al. 2021). However,

46 LSTM training time is long, the structure of the LSTM network is complex, and its parameters are

47 difficult to determine. Compared with the LSTM network, the gated recurrent unit (GRU) network

48 developed in recent years has a simpler structure and reduces computational complexity, so it has also

49 been applied in time series prediction (Jung et al. 2021; Pu et al. 2023; Li et al. 2023).

50 This paper focuses on a hybrid model of adaptive VMD and the GRU network for short-term load

51 forecasting. To reduce aliasing and random noise, this model uses an adaptive VMD method, determining

52 the modal number with the average sample entropy and mutual correlation. This model also uses the GRU

53 network, and further reduces training time by expanding the random adjustment parameters of the Adam

54 algorithm. The electric load forecasting results of this proposed model were compared with other state of

55 the art forecasting methods and found to be both accurate and reliable.

Adaptive VMD for load series

57 This section first introduces the basic principles of VMD, and then proposes an adaptive VMD

58 model to reduce aliasing and random noise in the load series.

1) Variational mode decomposition

60 VMD decomposes the original series into several intrinsic mode functions (IMFs), each of which is

61 a sub-sequence of the frequency modulation and amplitude modulation (Dragomiretskiy & Zosso 2014;

62 Zhang & Guo 2020). VMD demodulates the IMF to its own fundamental frequency bandwidth and aims

63   to minimize total modal bandwidth to find the optimal IMF. The VMD includes both variational

64   construction and a variational solution.

65   To obtain the analytical signals of each IMF and the corresponding unilateral spectrum, The modal

66   function obtained by decomposition, represented as: $u_k(t), k = 1,2,\text{L}, K$ is processed with Hilbert

67   transform (Huang et al. 1998) as follows:

68
$$\left(\delta(t) + \frac{j}{\pi t}\right) * u_k(t) \tag{1}$$

69   where $\delta(t)$ is the unit impulse signal.

70   Center frequency, represented as ω_k, is then multiplied by the exponential term $e^{-j\omega_k t}$ to modulate

71   the spectrum of the mode to its fundamental frequency:

72
$$\left[\left(\delta(t) + \frac{j}{\pi t}\right) * u_k(t)\right] e^{-j\omega_k t} \tag{2}$$

73   The bandwidth of the mode is then calculated, using the solution of the 2-Norm of the modulated

74   signal gradient to solve the variational constraint problem:

75
$$\min_{\{u_k\},\{\omega_k\}} \left\{\sum_{k=1}^{K} \left\|\partial_t\left[\left(\delta(t) + \frac{j}{\pi t}\right)u_k(t)\right]e^{-j\omega_k t}\right\|_2^2\right\}$$
$$s.t. \sum_{k=1}^{K} u_k(t) = f(t) \tag{3}$$

76   where $f(t)$ is the original signal.

77   The variational constraint problem is then transformed into an unconstrained problem using the

78   Lagrange multiplier method and quadratic multiplication operator alternation algorithm. By introducing

79   the Lagrange multiplication operator $\theta(t)$ and penalty factor C into the constraint problem, the

80   unconstrained problem is as follows:

81
$$L(\{u_k\},\{\omega_k\},\theta) = C\sum_{k=1}^{K}\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)u_k(t)\right]e^{-j\omega_k t}\right\|_2^2 + \left\|f(t) - \sum_{k=1}^{K}u_k(t)\right\|_2^2 + \left\langle\theta(t), f(t) - \sum_{k=1}^{K}u_k(t)\right\rangle \tag{4}$$

82   Then, the optimization problem of u_k can be obtained by using the multiplication operator alternating

83   algorithm:

84
$$u_k^{n+1} = \arg\min_{u_k \in X} \left\{\theta\left\|\partial_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)u_k(t)\right]e^{-j\omega_k t}\right\|_2^2 + \left\|f(t) - \sum_{k=1}^{K}u_k(t)\right\|_2^2 + \left\|f(t) - \sum_{i=1}^{K}u_i(t) + \frac{\theta(t)}{2}\right\|_2^2\right\} \tag{5}$$

85   where i is the iteration control parameter. By using the Parseval/Plancherel Fourier equidistant method

86   under 2-norm, it can be obtained as follows:

87
$$\hat{u}_k^{n+1} = \underset{\hat{u}_k, \, u_k \in X}{\arg\min} \left\{ \int_0^\infty 4\theta(\omega - \omega_k)^2 \left| \hat{u}_k(\omega) \right|^2 + 2 \left| \hat{f}(\omega) - \sum_{i=1}^K \hat{u}_i(\omega) + \frac{\hat{\theta}(\omega)}{2} \right|^2 d\omega \right\}$$ (6)

88   Therefore, the optimized solution for this quadratic problem is:

89
$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k, i=1}^K \hat{u}_i(\omega) + \frac{\hat{\theta}(\omega)}{2}}{1 + 2C(\omega - \omega_k)^2}$$ (7)

90   Finally, the center frequency can calculated using the following quadratic formula:

91
$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \left| \hat{u}_k(\omega) \right|^2 d\omega}{\int_0^\infty \left| \hat{u}_k(\omega) \right|^2 d\omega}$$ (8)

92   ## 2) Adaptive VMD

93   As a non-recursive decomposition method, the number of modes in VMD needs to be set in advance.

94   When the number of modes is too small, there is insufficient decomposition; also, some modal functions

95   can result in false spectrum and spectrum breakage (Jiang, Shen & Shi 2018). In our previous work, the

96   number of modes is set according to the cross-correlation coefficient (Shang, Li & Wu 2023). In the

97   proposed adaptive VMD model, to improve reliability the number of modes is jointly determined by the

98   cross-correlation coefficient and the average sample entropy.

99   **a) The correlation coefficient**

100   The correlation coefficient reveals the correlation degree of two sequences; the higher the correlation

101   between two sequences, the closer the cross-correlation coefficient is to 1 (Shang, Li & Wu 2023). With

102   the residual sequence represented as $l(n)$ and the modal function represented as $f(n)$, the standard cross-

103   correlation coefficient ρ_c is defined as:

104
$$\rho_c = \sum_{n=0}^N f(n)l(n) \Big/ \sqrt{R_{ff} R_{ll}}$$
   (9)

105   where N is the sequence length, and R_{ll} and R_{ff} are the autocorrelation coefficients of $l(n)$ and $f(n)$,

106   respectively.

107   **b) The average sample entropy**

108   Each mode after VMD decomposition has its own central frequencies, so the spectrum will not

109   overlap. As a result, the similarity of each mode is high, and the sample entropy is small. When the

110   number of decompositions is the optimal value, the sample entropy of each mode (except the residuals)

111   and the average sample entropy (ASE) should both be the smallest (Lake 2010; Sun & Wang 2018).

112   With the time series represented as $Y(n), n = 1,2,L\ , N$, and the modal function represented as

113    $u_k(n), k = 1,2,L\ , K$, $U(i)$ is obtained by the modal function of $u_k(i)$, as follows:

114
$$U(i) = \left[u_k(i), u_k(i+1), K\ , u_k(i+m-1) \right]$$
(10)

115   where $i = 1,2,L\ , N-m+1$.

116   Firstly, the maximum distance $d_m\left[U(i),\ U(j)\right]$ between the corresponding elements of $U(i)$ and $U(j)$

117   is defined as:

118
$$d_m\left[U(i),\ U(j)\right] = \max_{l=0,1,L\ ,m-1} \left| u_k(i+l) - u_k(j+l) \right|$$
(11)

119   Then, counting the number of j satisfying the formula $d_m\left[U(i),\ U(j)\right] < r$ for each i, defined as B_i.

120   Here, j satisfies $N-m \ge j \ge 1$ and r is the tolerance of similarity measure. Based on this, the ratio of B_i^m

121   to the total distance of $N-m$ is as follows:

122
$$B_i^m(r) = \frac{B_i}{N-m}$$
(12)

123   Next, the average value $B^m(r)$ of $B_i^m(r)$ is calculated as follows:

124
$$B^m(r) = \sum_{i=1}^{N-m} \frac{B_i^m(r)}{N-m+1}$$
(13)

125   Lastly, by increasing the dimension to $m+1$, the average value of B_i^{m+1} is obtained as follows:

126
$$B^{m+1}(r) = \sum_{i=1}^{N-m} \frac{B_i^{m+1}(r)}{N-m}$$
(14)

127   where $B^m(r)$ and $B^{m+1}(r)$ are the probability that two sequences match m and $m+1$ points under the

128   tolerance of similarity measure r, respectively. The sample entropy of the modal sequence can then be

129   written as:

130
$$S_E = -\ln\left[\frac{B^{m+1}(r)}{B^m(r)} \right]$$
(15)

131   Eq. (15) shows that sample entropy is related to both m and r. eference previous study (Pincus 2001)

132   showed that when r is 1 or 2 and m is $0.1 \sim 0.25$ STD (STD is the variance of the sequence), the sample

133   entropy is rarely affected by m and r. Therefore, m was set to 2 and r was set to 0.2 STD in this study.

134   In VMD decomposition of a time series, the components of different scales need to be separated so

135   they occupy their own spectrum bandwidth, and the random noise in the time series needs to be

136   distinguished from the modal components. Therefore, in this study, the average sample entropy and the

137   cross-correlation coefficient ρ_c were used to jointly determine the number of modes.

138   **Hybrid model for forecasting**

139   **1) GRU network**

140   The GRU network is a type of recurrent neural network (RNN; Cho et al. 2014) that has been

141   proposed to solve the problems of long-term memory in back propagation. The GRU network performs in

142   a similar way to the LSTM network but is computationally cheaper. The structure of a GRU network is

143   shown in Fig. 1 where x_t is the input at the current node t, h_{t-1} is the hidden state of transmission at the

144   previous node t-1, y_t is the output, and h_t is the hidden state at t.

145   The GRU network has two gate states, as shown in Fig. 2. Here, r and z are the reset gate and the

146   update gate, respectively, and can be written as:

$$r=\sigma\left(W_r\left[h_{t-1},x_t\right]+b_r\right) \tag{16}$$

$$z=\sigma\left(W_z\left[h_{t-1},x_t\right]+b_z\right) \tag{17}$$

149   where σ is the sigmoid function which constrains the data in the interval [0,1]; W and b are the weight and

150   threshold of the networks.

151   According to the reset gate r and the hidden state of h_{t-1}, the reset signal can be obtained, as follows:

$$h_{t-1}^{'}=h_{t-1}\otimes r \tag{18}$$

153   where $\otimes$ is the Hamiltonian operator.

154   The transmission state $h_t^{'}$ is written as:

$$h_t^{'}=\tanh\left(W_g\left[h_{t-1}^{'},x_t\right]+b_g\right) \tag{19}$$

156   Lastly, the transmission state of h_t can be written as:

$$h_t=(1-z)\otimes h_{t-1}+z\otimes h' \tag{20}$$

158   **2) The improved optimization algorithm**

159   The Adam algorithm updates weights by calculating the first and second moments of the gradient,

160   improving the slow convergence problem caused by the fixed learning rate in the gradient descent

161   method. The Adam algorithm was used on the random adjustment parameters in this study to effectively

162   improve the convergence rate.

163   Firstly, initializing the learning rate μ, and using optimization parameters W_t, the gradient g_t, the first

164   moment m_t of the gradient and the second moment v_t of the gradient can be calculated, iteratively, as

165   follows:

$$g_t=\nabla W_t f(W_t) \tag{21}$$

167 $$m_t = \beta_1 m_t + (1 - \beta_1) g_t \qquad (22)$$

168 $$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \qquad (23)$$

169 where β_1 and β_2 are the decay rates of the first and the second moments, respectively.

170   Then, the deviations of the first and second moments of the gradient can be calculated as:

171 $$m_t^{'} = m_t / (1 - \beta_1) - \eta_1 g_t \qquad (24)$$

172 $$v_t^{'} = v_t / (1 - \beta_2) - \eta_2 g_t^2 \qquad (25)$$

173 where η_1 and η_2 are the adjustment parameters of the first and second moments and the random numbers

174 on interval [0,1], respectively.

175   Lastly, the parameters can be updated using the following formula:

176 $$w_t = w_{t-1} - \mu m_t^{'} / \left(\sqrt{v_t^{'}} + \varepsilon \right) \qquad (26)$$

177 where ε is the allowable error to prevent a zero value in the iterative process.

178   **3) Forecasting process**

179   In this work, a hybrid model of adaptive VMD and the GRU network is proposed to reduce

180 frequency aliasing and eliminate the randomness of the load series. The electrical load series is

181 represented as $\{x(1), x(2), \mathrm{L}, x(N)\}$, where N is the number of samples. When C_i is the decomposed mode,

182 the prototype mode is calculated, as follows:

183 $$C_i = \{c_i(1), c_i(2), \mathrm{L}, c_i(N)\}, i = 1, 2, \mathrm{L}, M \qquad (27)$$

184 The forecasting value of the prototype mode at time $(N+1)$ is $\hat{c}_i(N+1)$ using the GRU networks.

185 Reconstructing other components after removing the residual sequence, the forecasting result at time $N+1$

186 is calculated, as follows:

187 $$\hat{x}(N+1) = \sum_{i=2}^{M} \hat{c}_i(N+1) \qquad (28)$$

188   Fig. 3 illustrates the load forecasting process of the hybrid model. First, the load series is

189 decomposed using the adaptive VMD model. Then, each mode is forecasted in the next time using the

190 improved GRU model. Finally, the sum of the IMF is the forecasting result of the load series. The

191 forecasting process is as follows:

192   a) The IMF of each modal component is obtained by initializing the number of modes and

193 decomposing the load data using the adaptive VMD model.

194   b) The cross-correlation coefficient ρ_c between the noise residue and IMF is calculated, using Eq.
195   (11).

196   c) The sample entropy of each mode and the modal residue, and the average sample entropy (ASE)
197   are calculated using Eqs. (9) to (15) under different numbers of modes.

198   d) The decomposition number K is selected, corresponding to the minimum of ASE and ρ_c.

199   e) The modal components at the next time point are then forecasted using the improved GRU
200   network.

201   f) The final result is obtained by reconstructing the modal forecasting components.

Results

203   To analyze the forecasting performance, the root mean square error (*RMSE*), the mean absolute error
204   (*MAE*) and the mean absolute percentage error (*MAPE*) were calculated, as follows:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(p_n - \hat{p}_n)^2}{N}} \tag{29}$$

$$MAE = \frac{\sum_{n=1}^{N}|p_n - \hat{p}_n|}{N} \tag{30}$$

$$MAPE = \frac{\sum_{n=1}^{N}\left|\frac{p_n - \hat{p}_n}{p_n}\right|}{N} \times 100\% \tag{31}$$

208   where p_n and $\hat{p}_n$ represent the actual value and the forecasted value at the time n, respectively, and N is
209   the number of samples.

210   As a case study, experimental electrical data were taken from the Shaanxi province, China. The
211   original load series was a one-minute interval, with data extracted every 15 minutes to form the data set.
212   Because of the weekly periodicity of the load series, the final set of data included 672 samples.

1) Results of the adaptive VMD model

214   The ASE of each component, except the residual and cross-correlation coefficient ρ_c, were
215   calculated, and then the minimum values of the mode number K were identified corresponding to the
216   minimum *ASE* and ρ_c. The minimum *ASE* indicates that the similarity of each IMF was high, and that the
217   sequence was more "orderly." The minimum ρ_c means that the correlation between the residual sequence
218   and the modal sequence was the smallest, meaning the reconstructed mode, except the residual, is closer
219   to the real load sequence. Fig. 4 shows the ASE and the cross-correlation coefficient ρ_c at different time

220   points; ASE reached the minimum value when the number of decompositions was 5, and ρ_c was the

221   smallest when $K=5$.

222   The proposed adaptive VMD model decomposes the original electrical load series. Fig. 5 shows the

223   decomposition results of the load series of a day, consisting of 96 samples. In this figure, mode IMF4 has

224   a small amplitude and violent random noise, giving it a residual sequence. The frequencies of the other

225   modes decrease from top to bottom, revealing the short-term and long-term characteristics of the load

226   series.

227   ## 2) Results of the improved GRU network

228   A training set was created from 100 groups of load series, with the other load series used as the test

229   set. The default values of related parameters are given in Table 1.

230   The forecasting performance of the improved GRU network was compared with the original GRU

231   model, with both methods using the same training and test data. Table 2 shows the comparison results of

232   both forecasting error and time. Each error index was based on the forecasting value of 672 datapoints.

233   There were 100 groups of error index values in total, and the average value was the error result shown in

234   Table 2. The forecasting error value of the improved GRU model proposed in this paper was less than the

235   forecasting error value of the original GRU model. The forecasting time of the improved GRU model was

236   also significantly reduced compared to the original GRU model because the Adam algorithm uses a

237   random adjustment of parameters.

238   Table 3 shows the forecasting error for different numbers of modes. When the number of modes was

239   4, the forecasting error was the smallest, verifying the reliability of the adaptive VMD model.

240   ## 3) Results of the hybrid model

241   The proposed forecasting model was then verified using different types of measured data. Fig. 6(a)

242   and Fig. 7(a) show the actual valuables and forecasting results (original GRU model and our hybrid

243   model) on a working day and a non-working day, respectively. In Fig. 6(a), the RMSE of the original

244   GRU model was 335.7 MW and the RMSE of the proposed model was 334.5 MW. In Fig. 7(a),

245   the RMSE of the original GRU model and the proposed model were 335.9 MW and 334.6 MW,

246   respectively. Since these differences were small, Fig 6(b) and Fig. 7(b) further illustrate the comparisons

247   with smaller units on the load (y-axis) and less forecasting points (x-axis). These figures show that the

248   proposed hybrid model had better forecasting performance than the original GRU model.

249   The number of hidden neurons is an important parameter that can affect forecasting performance.

250   Table 4 shows the forecasting results of different numbers of neurons of the hidden layer, with all other

251 parameters being optimal. The forecasting error was the smallest when the number of neurons was 40, but

252 the optimal number of neurons was different with different datasets.

253   The proposed hybrid forecasting model was also compared with the following classical statistical

254 models: the ARIMA model, support vector regression (SVR), machine learning (Elman neural network),

255 and the combined model. The parameter settings of the ARIMA model were set based on Lee & Ko

256 (2021), and the selection of model order was based on the AIC criterion. The kernel function of the SVR

257 model was the Gaussian radial basis function and the kernel parameters were optimized based on Sina &

258 Kaur (2020). The parameter settings of the Elman neural network were set according to Xie et al. (2020),

259 the optimization algorithm adopted the traditional gradient descent method, and the number of neurons in

260 both hidden layers was 40. The single method selection and parameter settings of the combined model

261 were based on Li & Chang (2018).

262   Table 5 illustrates the comparison of forecasting error of the above methods. Compared with the

263 traditional statistical model and machine learning, the hybrid model proposed in this work had a higher

264 forecasting accuracy. The traditional statistical learning method only obtained the evolution

265 characteristics of the time series based on a limited number of samples, making it difficult to forecast

266 long-term evolution and reversal characteristics of a time series. The machine learning methods, such as

267 the radial basis function (RBF) and back-propagation (BP) neural network, have poor forecasting ability

268 of a time series, and the Elman network is unable to forecast the long-term dependence of a time series.

269 The proposed hybrid model, based on the GRU network, is a deep learning method, which obtains the

270 evolution characteristics of sequences based on a large number of data, so the forecasting accuracy is

271 higher. Because it is a deep learning method, the training time of the hybrid model based on the GRU

272 network is much longer than that of the traditional statistical model and machine learning method.

273   Finally, the forecasting performance of the hybrid model in this work was compared with three state

274 of the art methods: the LSTM network, the GRU network and the QRNN model. The LSTM algorithm

275 and parameter settings were based on Rafi et al. (2021). The selection and parameter value of the GRU

276 network were based on Shen et al. (2021). The setting and parameter values of the QRNN network were

277 based on Cannon (2011). Table 6 compares the forecasting performance of the above methods. The

278 average forecasting error value shows that, compared with the LSTM forecasting method combined with

279 EMD and VMD, the AVMD model proposed in this paper decomposed the sequence more accurately and

280 improved the accuracy of the GRU network forecasting results. The maximum and minimum errors also

281 verified the stability and reliability of the hybrid model. Compared with other deep learning networks, the

282 parallel structure of the hybrid model based on the GRU network significantly shortened the training time,

283 making this model suitable for short-term power load forecasting.

284 Because it is a deep learning method, the hybrid model requires a large number of training samples,

285 and therefore the training time is longer than that of machine learning methods. Training time could be

286 further reduced by reducing the number of training samples and batch size through a consideration of the

287 periodicity of the load series. Furthermore, based on actual data, a reasonable network structure could be

288 optimized, such as the number of network layers and nodes, without significantly reducing forecasting

289 accuracy.

Discussion

291 The results of this study show the effectiveness of the established hybrid GRU network with

292 adaptive VMD for forecasting electrical load. The number of modes in adaptive VMD decomposition is

293 determined using the average sample entropy and the cross-correlation coefficient, improving forecasting

294 performance. The adaptive VMD decomposition eliminated the randomness of the series, and better

295 reflected the time scale characteristics of every subsequence, improving load forecasting performance,

296 although it increased the training time of the model.

297 Furthermore, we clarify the research gaps filled in by the proposed forecasting model. Because load

298 series have multiple periods and are nonlinear, the proposed model can improve forecasting accuracy by

299 decomposing the load series into multiple sub-sequences. It is also difficult to determine the number of

300 modes in VMD, and the proposed adaptive VMD can both determine the number of modes and improve

301 the reliability of decomposition using average sample entropy and the cross-correlation coefficient.

302 Finally, the proposed hybrid model improves forecasting accuracy at the cost of increased computational

303 time.

304 The computational complexity of the proposed hybrid model is similar to that of LSTM, GRU, and

305 RNN. Although the hybrid model has a longer computational time than traditional machine learning

306 methods, its forecasting accuracy is significantly improved. The size of the network and the number of

307 training samples could be further reduced in practical applications based on data characteristics, reducing

308 computational complexity and the corresponding computational time, making the hybrid model

309 appropriate for practical short-term power load forecasting.

Conclusions

311 This study established a hybrid model of adaptive VMD and the GRU network and applied the

312 model to short-term electrical load forecasting. The developed adaptive VMD method determines the

313   modal number using average sample entropy and mutual correlation. The developed GRU network

314   reduces training time by adding the random adjustment parameters to the Adam algorithm. The hybrid

315   model reduces frequency aliasing and the randomness of the series, so its forecasted loads are close to the

316   actual load data.

317   Some statistical models and machine learning methods, including ARIMA, SVR, the Elman

318   networks, and the combined model, and some state of the art models including the LSTM method and the

319   QRNN model, were compared with our proposed hybrid model. The values of MAPE, MAE and RMSE

320   were reduced in comparison with the traditional statistical models. The training time of the hybrid model

321   was much smaller than that of the deep learning method, and the proposed hybrid model had better

322   performance in short-term load forecasting.

References

324   **Cannon AJ**. 2011. Quantile regression neural networks: Implementation in R and application to

325   precipitation downscaling. *Computers & Geosciences* **37**(9):1277-1284

326   **Cho K, Merrienboer BV, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y.**

327   2014. Learning phrase representations using RNN encoder-decoder for statistical machine

328   translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language

329   Processing*, 1724-1734

330   **Dragomiretskiy K, Zosso D.** 2014. Variational mode decomposition. *IEEE Transactions on

331   Signal Processing* **62**(3):531-544

332   **Guan C, Luh PB, Michel LD, Chi ZY.** 2021. Hybrid Kalman filters for very short-term load

333   forecasting and prediction interval estimation. *IEEE Transactions on Power Systems*

334   **28**(4):3806-3817

335   **Huang NE , Shen Z, Long SR, Shen Z, Long SR, Wu ML, Shih HH, Zheng Q, Yen NC, Tung

336   CC, Liu HH.** 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear

337   and non-stationary time series analysis. *Proceedings Mathematical Physical & Engineering

338   Sciences* **454**:903-995

339   **Jiang X X, Shen C Q, Shi J J.** 2018. Initial centerfrequency-guided VMD for fault diagnosis of

340   rotating machines. *Journal of Sound and Vibration* **435**:36-55

341   **Jin ZY, Chakrabarti S, Yu J, Ding L, Terzija V.** 2021. An improved algorithm for cubature

342   Kalman filter based forecasting-aided state estimation and anomaly detection. *International

343   Transactions on Electrical Energy Systems* **31**(5):e12714

344   **Jung S, Moon J, Park S, Hwang E.** 2021. An attention-based multilayer GRU model for
345   multistep-ahead short-term load forecasting. *Sensors* **21**(5):1639

346   **Lake D.** 2010. Continuous sample entropy analysis. *Journal of Critical Care* **25**(3):e7-e8

347   **Lee CM, Ko CN.** 2021. Short-term load forecasting using lifting scheme and ARIMA models.
348   *Expert Systems with Applications* **38**(5):5902-5911

349   **Li C.** 2020. Designing a short-term load forecasting model in the urban smart grid system.
350   *Applied Energy* **266**:114850

351   **Li C, Li GJ, Wang KY, Han B.** 2023. A multi-energy load forecasting method based on parallel
352   architecture CNN-GRU and transfer learning for data deficient integrated energy systems.
353   *Energy* **259**:124967

354   **Ren J, Li WQ.** 2023. A robust maximum correntropy forecasting model for time series with
355   outliers. *Peer J Computer Science* **9**:e1251

356   **Li WQ, Chang L.** 2018. A combination model with variable weight optimization for short term
357   electrical load forecasting. *Energy* **164**:575-593

358   **Mideksa TK, Kallbekken S**. 2010. The impact of climate change on the electricity market: a
359   review. *Energy Policy* **38**(7):3579-3585

360   **Mokarram MJ, Rashiditabar R, Gitizadeh M, Aghaei J.** 2023. Net-load forecasting of
361   renewable energy systems using multi-input LSTM fuzzy and discrete wavelet transform.
362   *Energy* **275**:127425

363   **Mounir Nada, Ouadi H, Jrhilifa I.** 2023. Short-term electric load forecasting using an EMD-BI-
364   LSTM approach for smart grid energy management system. *Energy and Buildings*
365   **288**:113022

366   **Nobrega JP, Oliveira ALI.** 2019. A sequential learning method with Kalman filter and extreme
367   learning machine for regression and time series forecasting. *Neurocomputing* **337**:235-250

368   **Pincus SM**. 2001. Assessing serial irregularity and its implications for health. *Annals of the New
369   York Academy of Sciences* **954**:245-267

370   **Pu XW, Xiao H, Wang JR, Pei W, Yang W, Zhang JJ.** 2023. A novel GRU-TCN network based
371   Interactive Behavior Learning of multi-energy Microgrid under incomplete information. *Energy
372   Reports* **9**:608-616

373   **Quilty J, Aadamowski J.** 2018. Addressing the incorrect usage of wavelet-based hydrological
374   and water resources forecasting models for real-world applications with best practices and a
375   new forecasting framework. *Journal of Hydrology* **563**:336-353

376   **Rafi SH, Nahid-Al-Masood, Deeba SR, Hossain E.** 2021. A short-term load forecasting
377   method using integrated CNN and LSTM network. *IEEE Access* **9**:32437-32448

378   **Rayi VK, Mishra SP, Naik J, Dash PK.** 2022. Adaptive VMD based optimized deep learning
379   mixed kernel ELM autoencoder for single and multistep wind power forecasting. *Energy*
380   **244:**122585
381   **Shang T, Li WQ, W L.** 2023. Regional forecasting of wind speed in large scale wind plants.
382   *International Journal of Green Energy* **20**(5): 484-496
383   **Shen YM, Ma YX, Deng SM, Huang CJ, Kuo PH.** 2021. An ensemble model based on deep
384   learning and data preprocessing for short-term electrical load forecasting. *Sustainability*
385   **13**(4):1694
386   **Sina A, Kaur D.** 2020. Short term load forecasting model based on kernel-support vector
387   regression with social spider optimization algorithm. *Journal of Electrical Engineering &*
388   *Technology* **15**(1):393-402
389   **Sun W, Wang YW.** 2018. Short-term wind speed forecasting based on fast ensemble empirical
390   mode decomposition, phase space reconstruction, sample entropy and improved back-
391   propagation neural network. *Energy Conversion & Management* **157**:1-12
392   **Tarmanini C, Sarma N, Gezegin C, Ozgonenel O.** 2023. Short term load forecasting based on
393   ARIMA and ANN approaches. *Energy Reports* **9:**550-557
394   **Torres JF, Hadjout D, Sebaa A, Martinez-alvarez F, Troncoso A.** 2021. Deep learning for
395   time series forecasting: a survey. *Gig Data* **9**(1):3-21
396   **Wang Y, Wang J, Zhao G, Dong Y, Wang JZ.** 2012. Application of residual modification
397   approach in seasonal ARIMA for electricity demand forecasting: a case study of China.
398   *Energy Policy* **48**:284-294
399   **Xiao Z, Ye SJ, Zhong B, Sun CX.** 2007. BP neural network with rough set for short term load
400   forecasting. *Expert Systems with Applications* **36**:273-279
401   **Xie K, Yi H, Hu GY, Li LX, Fan ZY.** 2020. Short-term power load forecasting based on Elman
402   neural network with particle swarm optimization. *Neurocomputing* **416**:136-142
403   **Wang XL, Yao ZH, Papaethymiou M.** 2023. A real-time electrical load forecasting and
404   unsupervised anomaly detection framework. *Applied Energy* **330**:120279
405   **Zhang GQ, Guo J.** 2020. A novel ensemble method for hourly residential electricity
406   consumption forecasting by imaging time series. *Energy* **203**:117858

Default parameter settings of the improved GRU network.

Parameters	Default Valuables
Number of hidden layers	2
Number of Samples	100
Learning rate	0.01
Number of neurons in hidden layer 1	50
Number of neurons in hidden layer 2	50
Number of input sequences	672
Momentum parameter	0.5
Maximum number of iterations	2500
Optimization algorithm	Improved Adam algorithm

Comparison of forecasting performance of different GRU models.

Methods	RMSE(MW)	MAE(MW)	MAPE(%)	Training time(s)	Forecasting time(s)
GRU	336	201	1.931	323	3.59
Improved GRU	334	199	1.924	246	1.41

Forecasting error under different numbers of modes.

Number of Modes K	RMSE(MW)	MAE(MW)	MAPE(%)
2	347	217	2.103
3	340	214	2.012
4	**334**	**199**	**1.924**
5	337	210	2.006
6	342	220	2.118
7	347	218	2.095

Forecasting error under different numbers of neurons.

Number of neurons in hidden layer 1	Number of neurons in hidden layer 2	RMSE(MW)	MAE(MW)	MAPE(%)
10	10	359	208	2.101
20	20	354	205	1.999
30	30	338	197	1.912
40	**40**	**332**	**194**	**1.887**
50	50	339	198	1.986
60	60	345	203	2.002

Comparison of forecasting error of different forecasting methods.

Forecasting error		Forecasting method				
		ARIMA	SVR	Elman network	Combinational model	Our model
MAPE(%)	Averages	3.697	3.432	3.789	3.218	1.887
	Minimum	1.824	1.896	1.743	1.182	1.176
	Maximum	4.719	4.645	4.803	4.410	2.875
MAE (MW)	Averages	326	297	359	281	194
	Minimum	189	196	181	167	165
	Maximum	510	468	514	449	262
RMSE (MW)	Averages	475	437	507	421	332
	Minimum	316	328	303	280	277
	Maximum	773	731	778	682	395

Comparison of performance of deep learning methods.

Forecasting error		Forecasting method					
		EMD-LSTM	VMD-LSTM	EMD-GRU	VMD-GRU	QRNN	Our model
MAPE(%)	Averages	1.951	1.937	1.928	1.904	1.910	1.887
	Minimum	1.236	1.230	1.242	1.198	1.203	1.176
	Maximum	3.294	3.295	3.201	3.187	3.204	2.875
MAE (MW)	Averages	204	201	200	196	197	194
	Minimum	179	177	183	169	180	165
	Maximum	287	290	284	271	279	262
RMSE (MW)	Averages	348	342	340	335	338	332
	Minimum	291	288	299	280	283	277
	Maximum	413	418	408	397	402	395
Training time(s)	-	48.4	49.7	34.2	35.6	35.9	32.3

Figure 1

Input and output structure of the GRU network.
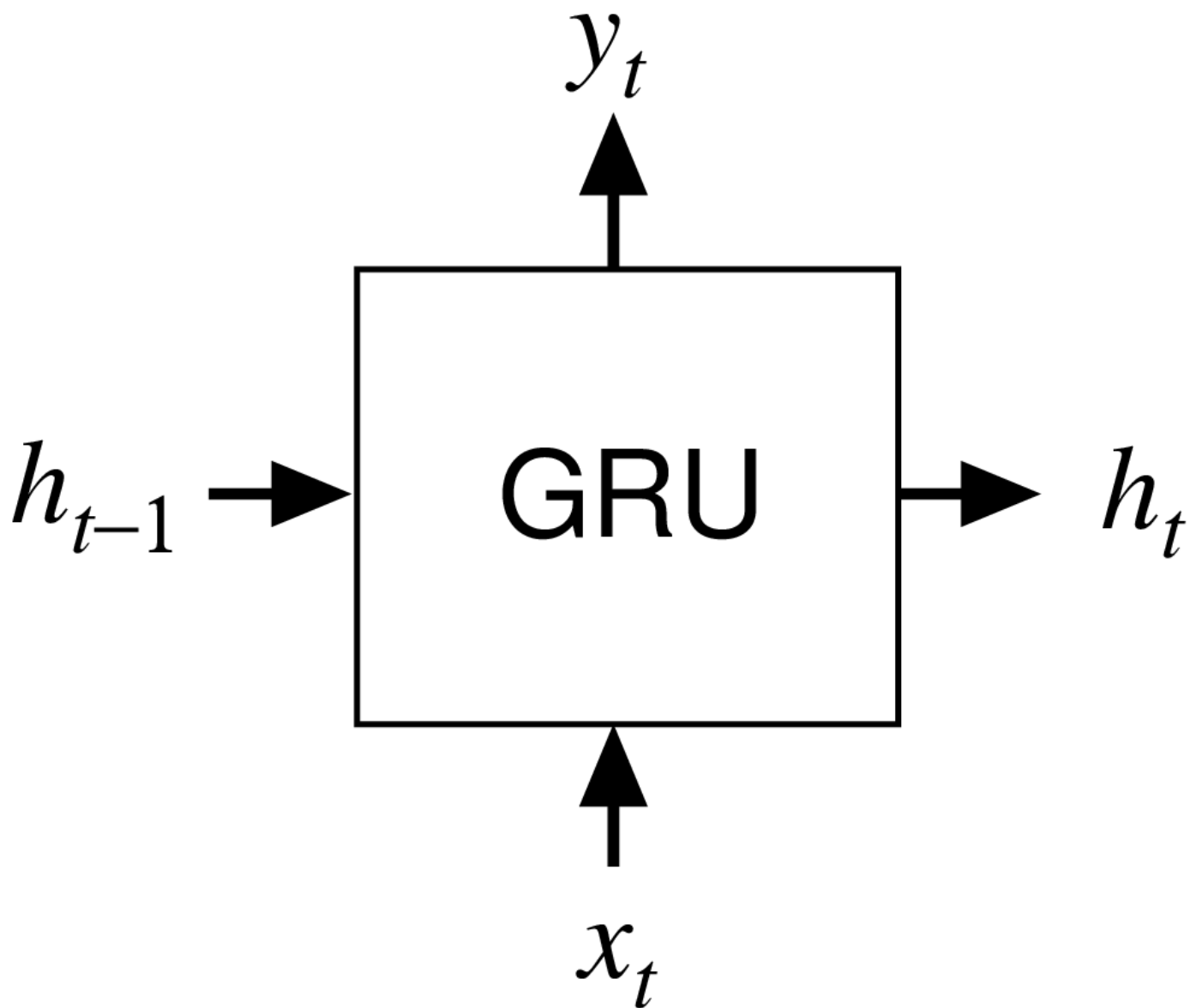
Figure 2

Gate structure of the GRU network.

$$r = \sigma\left(\boxed{W_r}\begin{bmatrix}\boxed{x_t}\\\boxed{h_{t-1}}\end{bmatrix}\right)$$

$$z = \sigma\left(\boxed{W_z}\begin{bmatrix}\boxed{x_t}\\\boxed{h_{t-1}}\end{bmatrix}\right)$$

Figure 3

Forecasting process of the hybrid model.
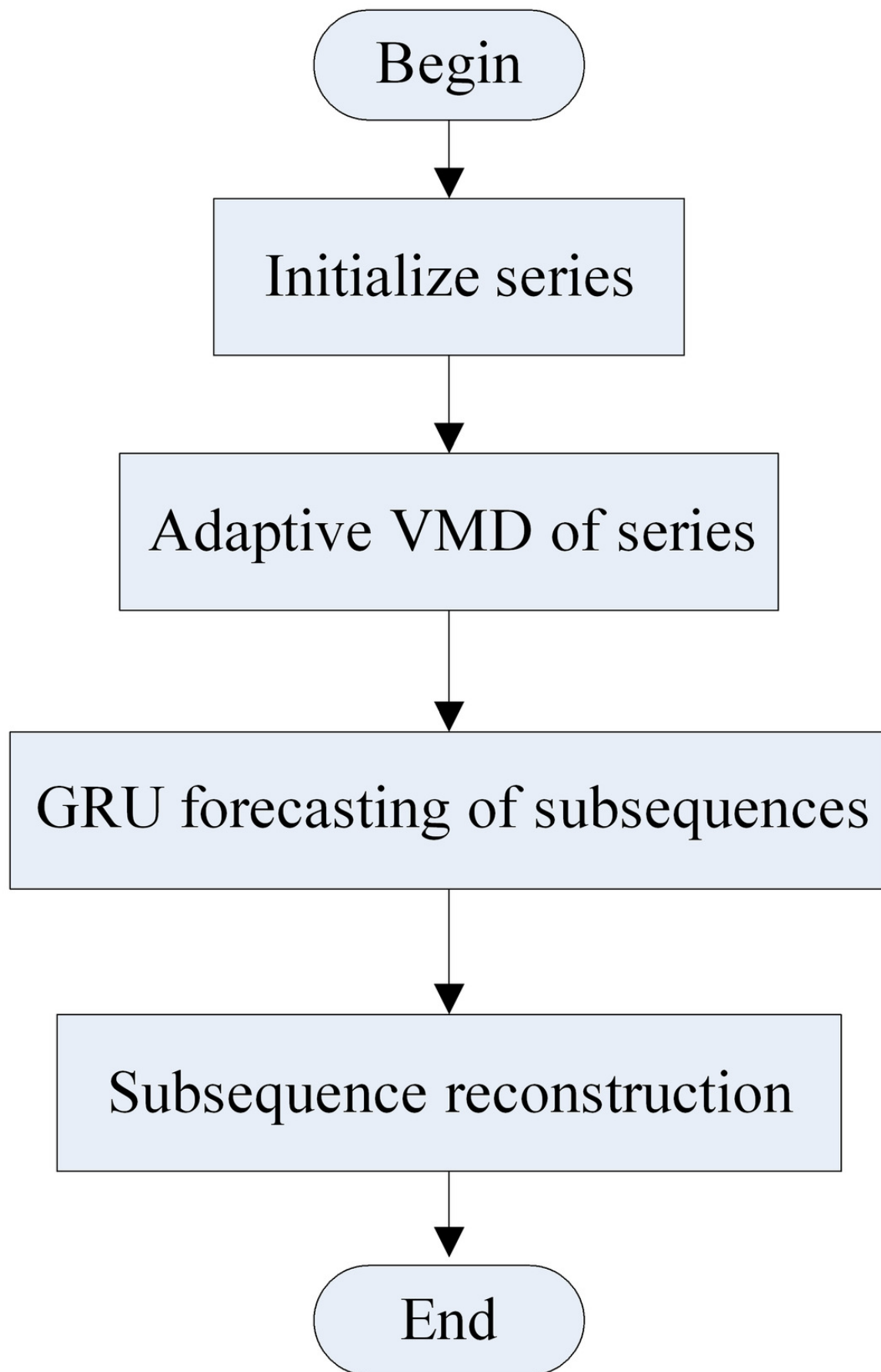
Figure 4

ASE and ρ_c of a load series of different K.
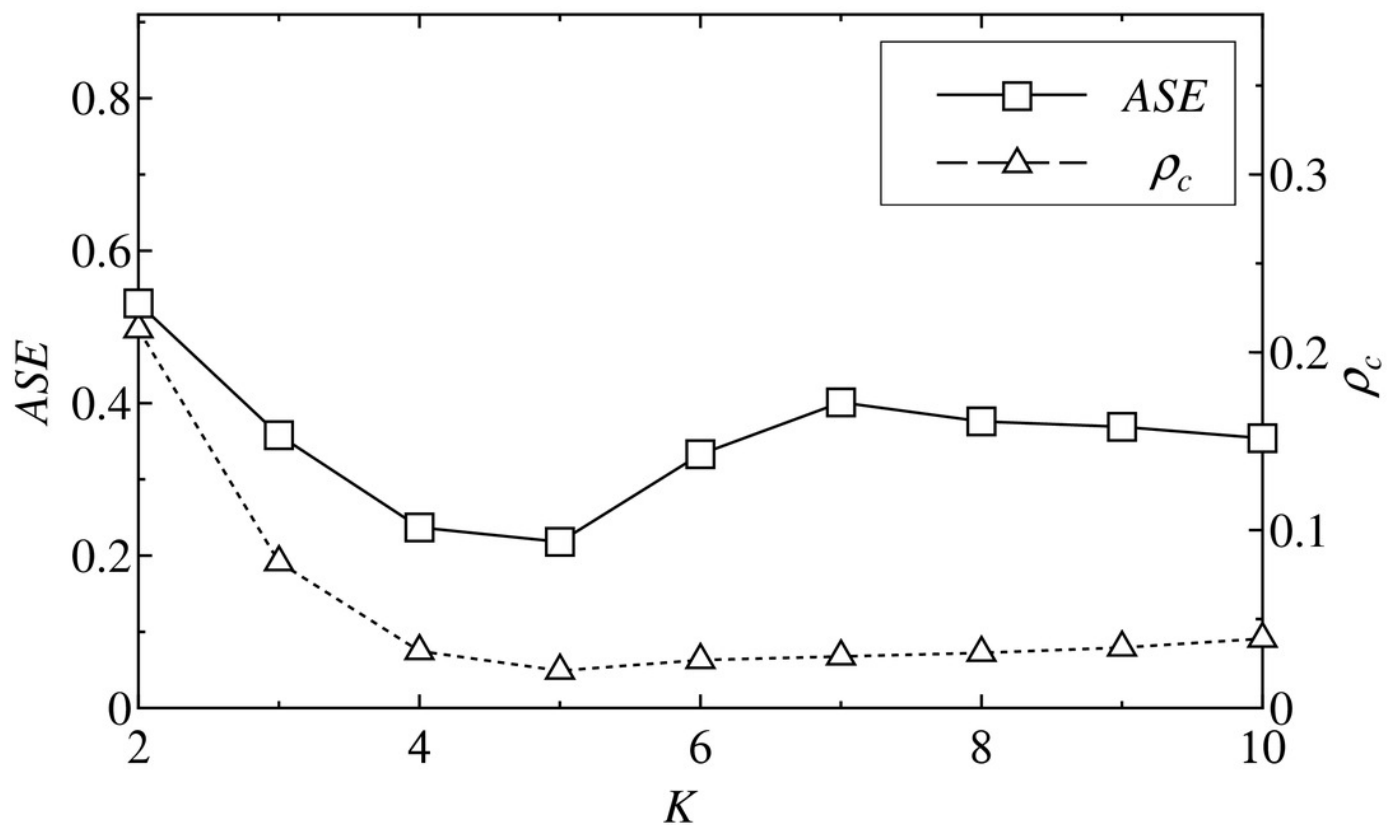
Figure 5

Electrical load series and decomposition results of the adaptive VMD.
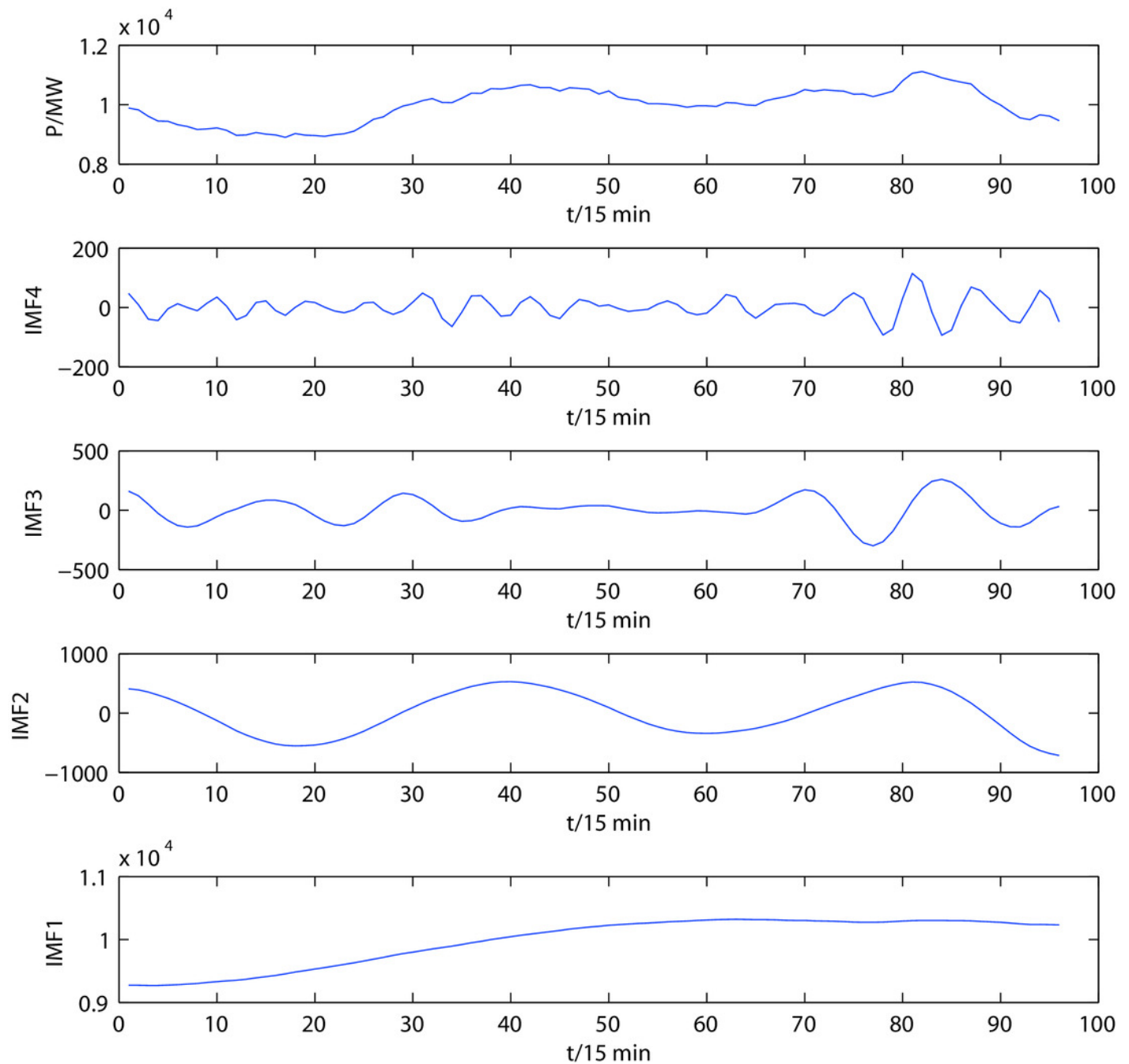
Figure 6

Forecast results of power loads on a working day.
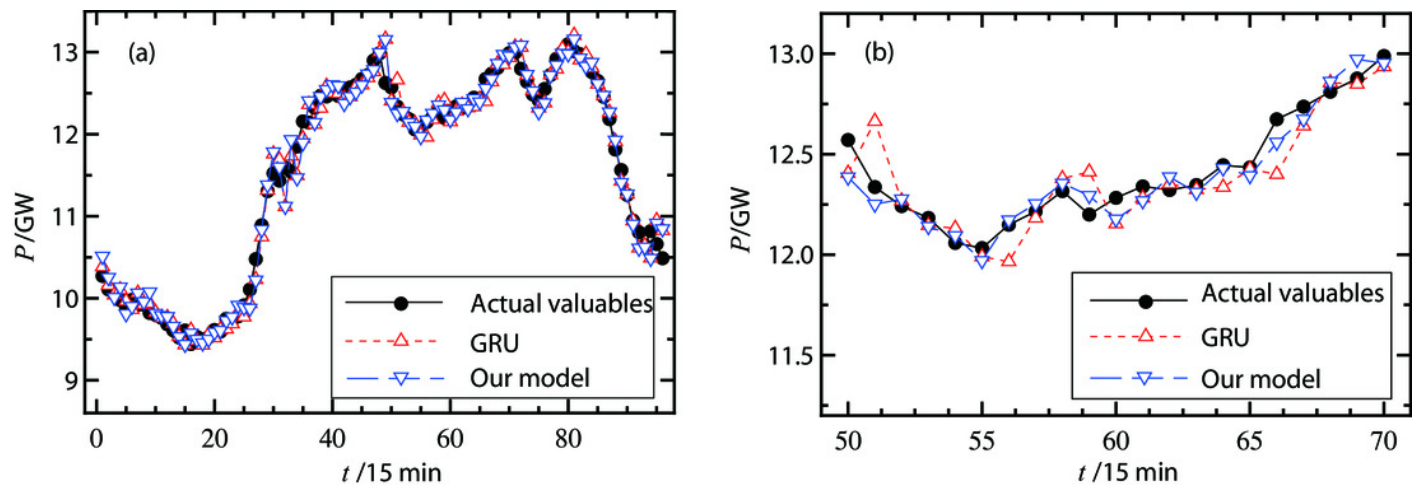
Figure 7

Forecast results of power loads on a non-working day.