ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

fETSmcs: Feature-based ETS model component selection

Lingzhi Qi [a,1], Xixi Li [b,*,1], Qiang Wang [a], Suling Jia [a]

[a] School of Economics and Management, Beihang University, China
[b] Department of Mathematics, University of Manchester, UK

ARTICLE INFO

Keywords:
ETS
Model component selection
Time series features
M4 dataset
Hospital dataset

ABSTRACT

The well-developed ETS (ExponenTial Smoothing, or Error, Trend, Seasonality) method incorporates a family of exponential smoothing models in state space representation and is widely used for automatic forecasting. The existing ETS method uses information criteria for model selection by choosing an optimal model with the smallest information criterion among all models fitted to a given time series. The ETS method under such a model selection scheme suffers from computational complexity when applied to large-scale time series data. To tackle this issue, we propose an efficient approach to ETS model selection by training classifiers on simulated data to predict appropriate model component forms for a given time series. We provide a simulation study to show the model selection ability of the proposed approach on simulated data. We evaluate our approach on the widely used M4 forecasting competition dataset in terms of both point forecasts and prediction intervals. To demonstrate the practical value of our method, we showcase the performance improvements from our approach on a monthly hospital dataset.

© 2022 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

1. Introduction

In today's data-rich environment, modern business decisions often involve forecasting large collections of time series data, and people often want to select optimal models for their data in an effective and automatic manner (Petropoulos et al., 2022). However, selecting one optimal model requires professional knowledge and relevant experience, making automatic and accurate forecasting not a straightforward task (Li, Petropoulos, & Kang, 2022).

Hyndman, Koehler, Snyder, and Grose (2002) incorporated a family of exponential smoothing models into a well-developed modeling framework for automatic forecasting, called ETS (ExponenTial Smoothing, or Error, Trend, Seasonality). The ETS method is able to identify three components of a given time series—error, trend, and seasonality, as well as their component forms (additive or multiplicative). The error component can be identified as additive ('A') or multiplicative ('M'). The trend and seasonality components can be none ('N'), 'A', or 'M'. Also, the trend can be damped or not.

In the context of ETS modeling, model selection refers to identifying appropriate forms of three components (error, trend, and seasonality) for a given time series. The ETS method uses information criteria such as the Akaike information criterion (AIC), the corrected AIC (AICc), and the Bayesian information criterion (BIC) for selection purposes, and it proceeds in two steps: (i) calculating the information criteria of all candidate models fitted to data, and (ii) selecting a model with the smallest information criterion among all models. However, in a modern business environment, large retailers often need to forecast large-scale data from thousands of products in a forecast cycle (Li et al., 2022). For example, Walmart has roughly 5,000 stores in the U.S., and it must generate approximately one billion unique forecasts for business

* Corresponding author.
E-mail addresses: lingzhi_qi@buaa.edu.cn (L. Qi), xixi.li@manchester.ac.uk (X. Li), wang6965@sina.com (Q. Wang), jiasuling@buaa.edu.cn (S. Jia).

[1] The authors contributed equally.

https://doi.org/10.1016/j.ijforecast.2022.06.004
0169-2070/© 2022 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

Please cite this article as: L. Qi, X. Li, Q. Wang et al., fETSmcs: Feature-based ETS model component selection. International Journal of Forecasting (2022), https://doi.org/10.1016/j.ijforecast.2022.06.004.

decisions (Seaman, 2018). The existing ETS method under such a model selection scheme is computationally expensive when used to forecast such large-scale time series data. There is an urgent need for an efficient approach to ETS model selection to forecast large collections of time series effectively.

The issue of the computational cost of a large family of models has attracted some attention in recent studies. Nikolopoulos and Petropoulos (2018) proposed suboptimal forecasting solutions for faster and robust systems. They compared the performance of a suboptimal search for smoothing parameters of exponential smoothing models with that of a more computationally expensive and optimal search. Petropoulos and Grushka-Cockayne (2021) also questioned the necessity of such a large family of models, and suggested a change from a focus on the search of suboptimal parameters to that of suboptimal models.

A data-driven approach that uses time series features for forecast model selection and combination over large-scale time series data has flourished in recent studies (e.g., Li, Kang, & Li, 2020; Montero-Manso, Athanasopoulos, Hyndman, & Talagala, 2020; Talagala, Li, & Kang, 2021). However, research on existing feature-based model selection focuses more on selection among different types of models. The use of features for model selection among a family of models has not yet attracted enough attention. In order to improve the efficiency of the ETS method, we explore the possibility of using features for ETS model selection by training classifiers on simulated data to predict model component forms for a given time series.

We show performance improvements in model selection from our method on simulated data. We conduct extensive empirical evaluations of the proposed method by applying it to 95,000 real time series from the Makridakis forecasting competitions (Makridakis, Spiliotis, & Assimakopoulos, 2020). Significance tests results show that the proposed approach results in robust forecasts at long-term and all horizons for yearly, quarterly, and monthly data. To demonstrate the practical value of the proposed approach, we show the performance improvements from our method on a hospital dataset compared with the information criteria method.

The key innovations of our approach are as follows.

- The existing use of features for model selection focuses more on selection among different types of models, while our work makes the first attempt to select an appropriate model among a family of models.
- The proposed ETS model selection approach can effectively reduce computational cost when applied to large-scale time series data by predicting model component forms using pre-trained classifiers.

The rest of the article is organized as follows: Section 2 offers a short literature review of relevant studies. Section 3 describes the methodology for the proposed approach. Section 4 presents the empirical results, including a simulation study, real data applications, and significance tests. Section 5 offers a case study. Section 6 presents computation time for producing forecasts. Section 7 offers our discussion and insights. Finally, Section 8 provides our conclusions.

2. Background research

2.1. The exponential smoothing family of models

Exponential smoothing methods have been around since the 1950s when Brown (1956) proposed a simple exponential smoothing method for demand prediction. The work of Brown (1959) and Gardner Jr. (1985) led to the development of exponential smoothing models towards automated forecasting (Hyndman et al., 2002). Ord, Koehler, and Snyder (1997) and Hyndman et al. (2002) showed that all exponential smoothing models can be well written in the form of state space representation. Hyndman et al. (2002) incorporated a family of exponential smoothing models into a well-developed modeling framework called ETS. The framework incorporates state space models, parameter estimation, point forecasts, and prediction intervals.

2.2. Information criteria for model selection

Information criteria approaches are designed for parametric methods under the framework of maximum likelihood estimation. They are used to correct for the bias of maximum likelihood by adding a penalty term to compensate for the over-fitting of more complex models (Bishop, 2006). These commonly used information criteria are employed to evaluate how well a given model fits the data, and how complicated the fitted model is.

Although information criteria partially address the problem of over-fitting (Meira, Oliveira, & Jeon, 2021), the fact that the information criteria are based on one-step prediction errors implies that the selected model does not necessarily produce accurate and reliable forecasts when applied to long forecasting horizons (Fildes & Petropoulos, 2015). To tackle this issue, cross-validation can be used to assess the performance of models over multiple time windows. But this is not necessarily guaranteed, and Billah, King, Snyder, and Koehler (2006) showed that the performance of information criteria approaches is superior to that of the commonly used validation approach. However, the use of cross-validation is suitable for longer series and is computationally expensive (Li et al., 2022). Moreover, information criteria approaches do not consider parameter uncertainty (Bishop, 2006), and in practice, they tend to select simple models.

2.3. Feature-based model selection and combination

The use of machine learning for model selection is not novel. Many attempts have been made at feature-based forecasting for univariate time series. Reid (1972) pointed out that forecasting performance is influenced by time series features. Kang, Hyndman, and Smith-Miles (2017) projected time series into an instance space and provided some insights into the forecasting performance of different models. Recently, time series features have been used for forecast model selection and combination, aiming at selecting an appropriate model or at predicting the combination weights of all candidate models for a given time series. The performance of feature-based

Table 1
Taxonomy of exponential smoothing models in the ETS framework (Hyndman et al., 2008). The 15 applicable models are shaded in purple. The 11 models that lead to infinite forecast variance problems are shaded in gray. The four models with multiplicative trends that lead to unrealistically explosive forecasts are shaded in pink.

Trend	Additive error			Multiplicative error		
	Seasonality			Seasonality		
	N	A	M	N	A	M
N (None)	ANN	ANA	ANM	MNN	MNA	MNM
A (Additive)	AAN	AAA	AAM	MAN	MAA	MAM
Ad (Additive damped)	AAdN	AAdA	AAdM	MAdN	MAdA	MAdM
M (Multiplicative)	AMN	AMA	AMM	MMN	MMA	MMM
Md (Multiplicative damped)	AMdN	AMdA	AMdM	MMdN	MMdA	MMdM

forecasting is highly dependent on the selection of an appropriate set of time series features. Collopy and Armstrong (1992) developed a rule-based system to select optimal models for the series based on 99 features. Talagala, Hyndman, Athanasopoulos, et al. (2018) trained a decision tree using 42 features for the purpose of model selection. Montero-Manso et al. (2020) trained XGBoost to predict an optimal combination weight of each candidate model. Their method ranked second at the M4 competition (Makridakis et al., 2020). Talagala et al. (2021) predicted forecasting errors using time series features with a Bayesian multivariate surface regression model. Instead of using manual features, Li et al. (2020) proposed an automated approach to extracting time series imaging features used for forecast model combination.

3. Feature-based ETS model component selection

3.1. Taxonomy of exponential smoothing models in the ETS method

Table 1 presents the taxonomy of exponential smoothing models in the ETS method. Theoretically, the number of exponential smoothing models from the combinations of three component forms can reach 30, as shown in Table 1. However, in practice, the 11 models illustrated in Table 1 with a gray background would lead to infinite forecast variance problems when applied to long forecasting horizons. For more details on this issue, see Chapter 15 of Hyndman, Koehler, Ord, and Snyder (2008). Considering that models with multiplicative trends would lead to unrealistically explosive forecasts (Petropoulos & Grushka-Cockayne, 2021), the R package **forecast** (Hyndman et al., 2020) has, by default, excluded these four models ('MMN', 'MMdN', 'MMM', and 'MMdM') by setting the argument `allow.multiplicative.trend` as `False` in the function `ets()`. Hence, there are 15 applicable models
(shaded in purple in Table 1) in the ETS method. For non-seasonal data (e.g., yearly data), there are six applicable models.

3.2. The proposed framework

The aim of this work is to employ machine learning methodology to predict an optimal ETS model for a given time series. A simple approach to this task would be to view it as a multi-label classification problem. That is, we train a classifier to link time series features with the 15 applicable ETS models in Table 1. However, training a classifier with 15 classes is not an easy task. To tackle this issue, an alternative way is to adopt a divide-and-conquer strategy (Zhou, 2021) that trains a classifier to predict each component form separately.

Fig. 1 shows the proposed framework of feature-based ETS model selection consisting of the following two parts:

1. **Offline (training phase).** The training phase consists of the process of data simulation and feature extraction, as well as building and evaluating classifiers. In the simulation process, we simulate some time series from the 15 applicable ETS models. The extracted features over the simulated series with corresponding model component forms are used to build and evaluate classifiers. Three classifiers are trained on simulated data separately for the prediction of error, trend and, seasonality component forms.
2. **Online (testing phase).** The testing part requires feature extraction over any new time series. These extracted features are used as input to the pre-trained classifiers from the training phase to obtain error, trend, and seasonality component forms. We check the applicability of a selected model from the combination of predicted component forms, and model adjustment is performed depending on whether the selected model is applicable. Finally, we fit the selected model to the data using the function `forecast::ets()` (Hyndman et al., 2020) with the argument `model` being the selected model form, and then make predictions from the fitted model using the function `forecast::forecast()` (Hyndman et al., 2020).

3.3. Feature extraction

A collection of 42 time series features developed by Montero-Manso et al. (2020) was used for forecast model selection and combination over large collections of time series with good results in recent studies (e.g., Montero-Manso et al., 2020; Talagala et al., 2021; Wang, Kang, Petropoulos, & Li, 2021). The combination method of Montero-Manso et al. (2020) ranked second in one of largest time series competitions, namely, the M4 competition (Makridakis et al., 2020). The collection of features

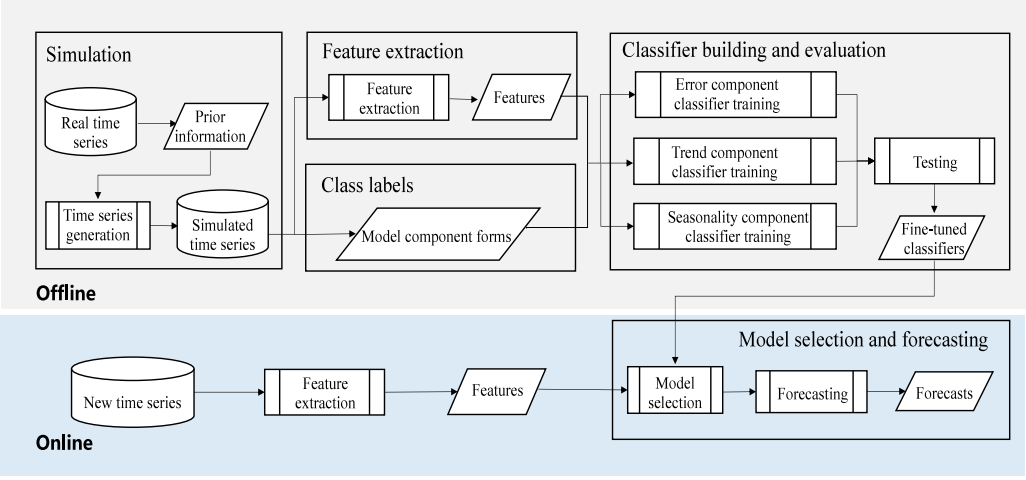

Fig. 1. The proposed framework of feature-based ETS model component selection.

is available in the R package **M4metalearning** (Montero-Manso et al., 2020).

Apart from the above features used for forecasting, a collection of 22 high-performing features carefully selected from 7000 features in the **hctsa** toolbox (Fulcher & Jones, 2017) exhibited strong classification performance across different time series problems (Lubba et al., 2019). This collection of features is available in the R package **Rcatch22** (Henderson, 2021). In this work, in order to characterize an observed time series more comprehensively, and thus improve the classification accuracy of model selection, we chose 59 features from these two classes of features for feature extraction. We discarded a set of five features from Montero-Manso et al. (2020) that use the estimated parameters of fitted ETS models to characterize a time series. The specific features and their corresponding descriptions can be found in Appendix A.

3.4. Classifier building for the prediction of ETS model component forms using LightGBM

XGBoost and pGBRT are effective implementations of a popular machine learning algorithm GBDT (Gradient Boosting Decision Tree) with many engineering optimizations. Nonetheless, the efficiency and scalability of both are still unsatisfactory when applied to data with high feature dimension and large size (Ke et al., 2017). To tackle this issue, Ke et al. (2017) proposed a new GBDT implementation called LightGBM with two novel techniques: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). With GOSS, only small data with large gradients are used to estimate the information gain. With EFB, the number of features is reduced by bundling mutually exclusive features.

In this study, instead of training a classifier with 15 classes using LightGBM, we separately train three classifiers f_e, f_t, and f_s to link time series features with each model component form. These three classifiers for the prediction of error, trend, and seasonality component forms are shown in Table 2.

3.5. Model check and adjustment

Each pre-trained classifier provides a certain probability that each component form is selected. Before fitting a selected model to data, we must confirm whether the model from the combination of three predicted component forms is applicable. Fig. 2 presents the specific procedure of model checking and the corresponding adjustment.

- **Check 1.** Because we train each classifier on all simulated data with different frequencies, it is unavoidable to mistakenly predict a seasonal ETS model for yearly data. If so, we set the seasonality component of the selected model as 'N'.
- **Check 2.** If a selected model is inapplicable (like 'ANM', 'AAM', and 'AAdM'), we replace its error and seasonality component forms with those from a model with the largest probability of being selected among all alternative models. The selected probability of an alternative model is calculated by multiplying the probabilities of each component.
- **Check 3.** Applying a selected model with multiplicative errors to a time series with zero or negative values would lead to the instability of the model (Hyndman & Athanasopoulos, 2018). To tackle this issue, we change its error component form to 'A'. Subsequently, we need to check whether the updated model is applicable. If not, we replace its seasonality component form with the one with the largest probability among all alternative forms.
- **Check 4.** To ensure a valid estimation of parameters in a model with a damped trend, the function `forecast::ets()` (Hyndman et al., 2020) requires that the number of time series should be larger than the number of parameters plus four when fitting it to data. If this condition is not met, we change a damped trend form to another one with a selected probability being ranked second among all trend component forms.

Table 2
Three classifiers used for the prediction of three component forms of an ETS model for a given time series.

Classifier	Description	Number of classes	Class labels
f_e	A classifier to predict the error component form	2	'A', 'M'
f_t	A classifier to predict the trend component form	3	'A', 'Ad', 'N'
f_s	A classifier to predict the seasonality component form	3	'A', 'M', 'N'

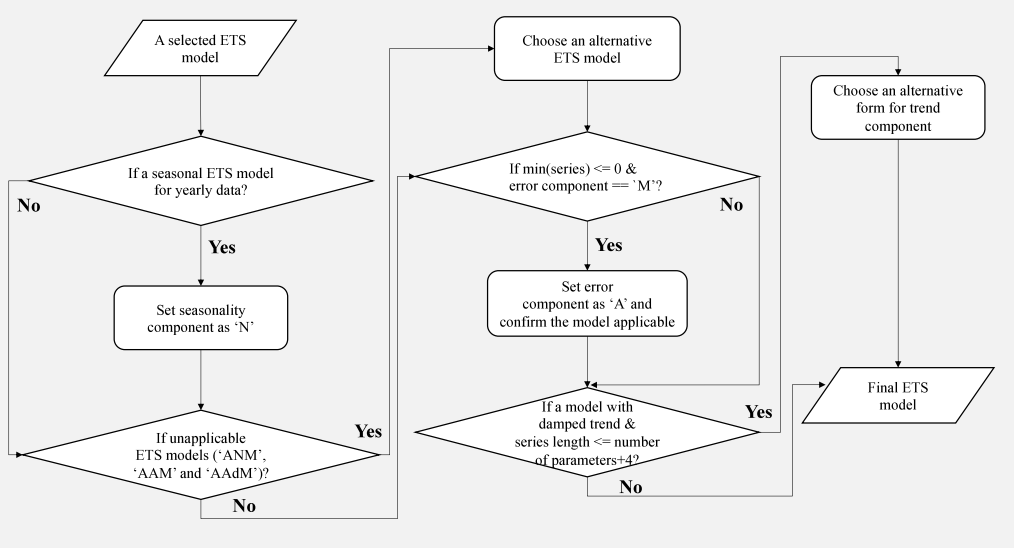

Fig. 2. The procedure of model checking and adjustment.

4. Empirical evaluation

4.1. Simulation study

4.1.1. Data simulation

Considering that the M4 competition data (Makridakis et al., 2020) are highly representative of reality (Spiliotis, Kouloumos, Assimakopoulos, & Makridakis, 2020), we obtained some prior information from the M4 data to simulate realistic series for the simulation study, as shown in Table 3. The length sets for simulating time series with different frequencies were obtained from the M4 dataset. Particularly, for an ETS model with 'Ad' trend, we set the *phi* parameter to control the degree of damping, and we let $phi \sim U(0.80, 0.98)$ with reference to the function `forecast::ets()` (Hyndman et al., 2020). A model set with 15 applicable ETS models was used for quarterly and monthly data simulation, and another one with six models was used for yearly data. In simulating a time series with a specific frequency, we randomly chose an ETS model form and length value from the corresponding model set and length set.

Finally, we simulated 60,000 yearly, 60,000 quarterly, and 120,000 monthly time series. The numbers for each seasonal and non-seasonal ETS model are approximately 12,000 and 22,000, respectively. Fig. 3 shows some typical simulated time series from the 15 applicable models.

t-SNE (Van der Maaten & Hinton, 2008) is a nonlinear dimensionality reduction technique used to project high-dimensional data into a low-dimensional space for visualization. To visualize whether the simulated dataset has a good representation of reality, we projected the simulated data and the M4 data into a two-dimensional feature space with t-SNE, as shown in Fig. 4. Given the two-dimensional feature spaces of the M4 and simulated datasets, we quantified the miscoverage of the simulated dataset over the M4 dataset in the following steps (Kang, Hyndman, & Li, 2020).

- Cut the x and y dimensions of the instance space into $N_b = 30$ bins, and obtain $N_b^2 = 900$ subgrids.
- In the constructed subgrids, let

$$\mathcal{I}_{i,\,\mathrm{S}} = \begin{cases} 0 & \text{if no time series in the simulated dataset} \\ & \text{fall into the } i\text{th subgrid;} \\ 1 & \text{otherwise.} \end{cases} \tag{1}$$

An analogous definition of $\mathcal{I}_{i,\,\mathrm{M}}$ is given for the M4 dataset.

Table 3
Parameter settings for time series simulation using `smooth::sim.es()` (Svetunkov, 2021).

Parameter	Description	Value
`frequency`	Period of time series	1: yearly; 4: quarterly; 12: monthly
`model`	The 15 applicable models, six for yearly data	Randomly chosen from the model sets
`obs`	Length of time series	Randomly chosen from the length sets obtained from the M4 dataset
`phi`	Damping parameter for a model with 'Ad' trend	U(0.80, 0.98)

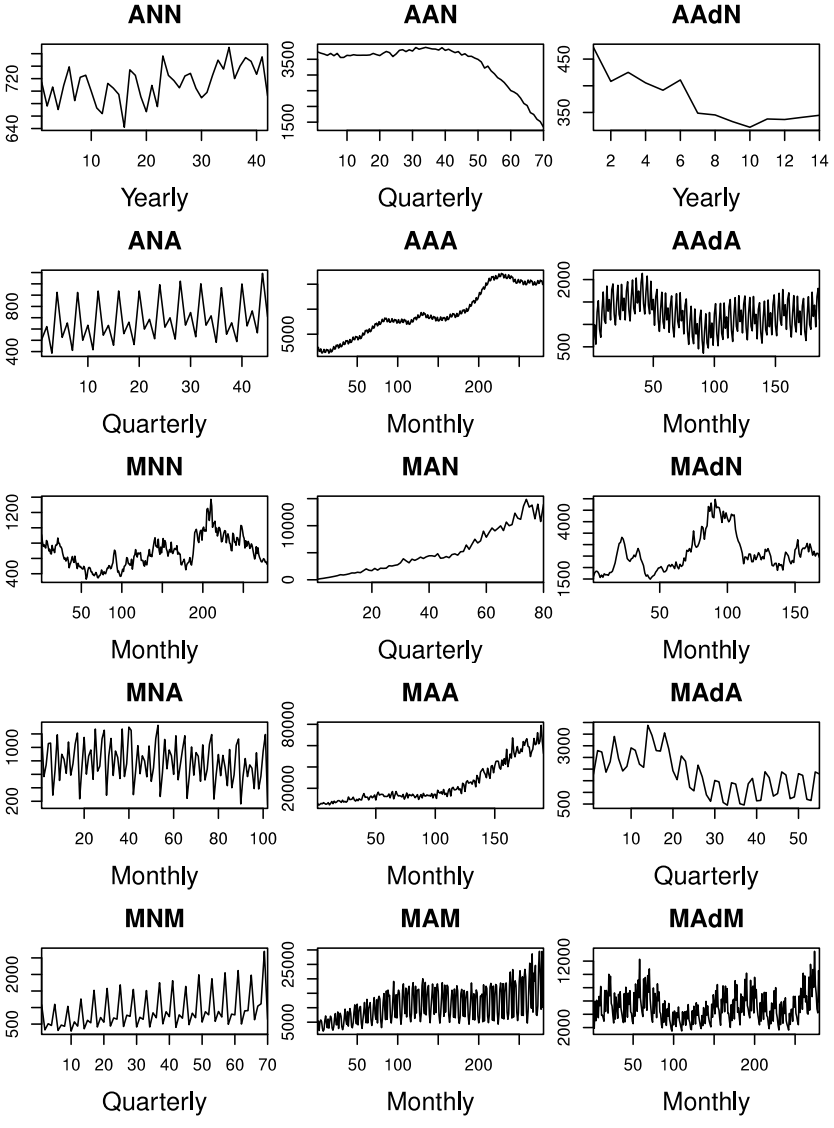

Fig. 3. Typical simulated series from the 15 models in the ETS method.

- The relative miscoverage of the simulated dataset over the M4 dataset is

$$\text{miscoverage}_{\text{S/M}} = N_b^{-2} \sum_{i=1}^{N_b^2} \left[\left(1 - \mathcal{I}_{i,\,\text{S}}\right) \times \mathcal{I}_{i,\,\text{M}} \right]. \quad (2)$$

The miscoverage values of the simulated data over the M4 data are 0.026, 0.003, and 0.002 for yearly, quarterly, and monthly data, respectively. Together with Fig. 4, we conclude that the simulated dataset has good coverage over the M4 data, and thus has good representativeness of reality.

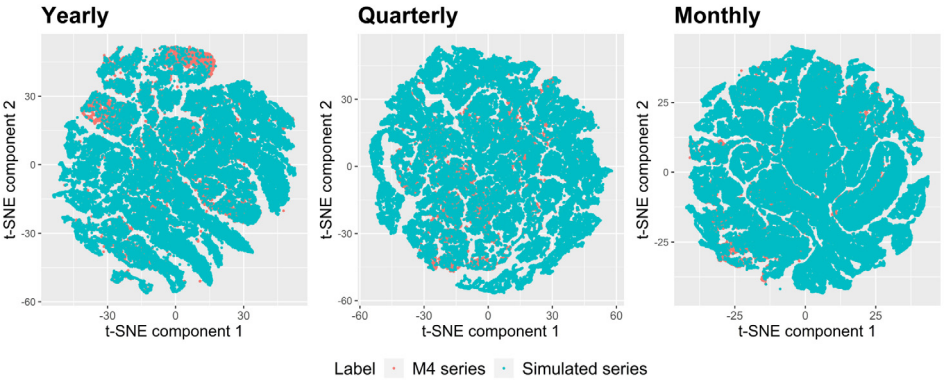

Fig. 4. M4 and simulated datasets in a two-dimensional feature space using t-SNE.

4.1.2. Performance of model selection on the simulated series

We split the simulated data into two parts: 80% of the series are used for training data, and the rest are used for testing data. Three sets of optimal hyperparameters of LightGBM from a search in subspaces of the hyperparameter spaces are obtained by using a five-fold cross-validation procedure on the training data. More details about these optimal hyperparameters are available in Appendix B.

In this simulation study, to demonstrate the model selection ability of the proposed method, we compare its performance on the testing data with that of the information criteria method. Two widely used metrics, the accuracy and macro F_1-score, are used to evaluate the classification. The accuracy is calculated as

$$\text{Accuracy} = \frac{TS}{N} \times 100\%, \tag{3}$$

where TS is the number of series with true selected component forms or models, and $N = 48{,}000$ is the number of testing series.

The F_1-score$_i$

$$F_1\text{-score}_i = 2\frac{Recall_i Precision_i}{Recall_i + Precision_i} \times 100\% \tag{4}$$

is used to measure the classification performance of the ith class by combining its recall and precision into a single metric.

The macro F_1-score

$$\text{Macro } F_1\text{-score} = (\frac{1}{m}\sum_{i=1}^{m} F_1\text{-score}_i) \times 100\% \tag{5}$$

gives the same importance to each class by equally averaging the F_1-score$_i$, $i = 1, \ldots, m$ of all classes, where m is the number of classes.

Table 4 presents the classification performance of our method against the information criteria method in the tasks of separate ETS model component selection and whole model selection. The ETS method using information criteria for model selection is implemented using the function `forecast::ets()` (Hyndman et al., 2020), with AICc as a default model selection criterion.

From Table 4, we can observe that our proposed method is superior to the information criteria method in all tasks on the simulated data.

4.1.3. Feature importance analysis for each component selection

For a better understanding of which features are important to each component form prediction, we identify the ten most important features among the 59 features by comparing their information gain (Ke et al., 2017). The higher the gain, the more important the feature. Fig. 5 presents the feature importance results for each component form prediction. We can observe the following.

- **Error component.** The error-related feature `spike` that calculates the variance of the leave-one-out variances of the remainder from an STL decomposition plays the most important role in the identification of the error component form. The feature `flat_spots` that measures the longest flat spot in a series also has a significant effect on the identification of the error component form. Further, the seasonal and trend strength of a series has an impact on error component selection in the sense that the error component is obtained after detrending and seasonal adjustment.
- **Trend component.** The trend-related features `trend` and `unitroot_pp` are identified as the two most important features. The feature `localSimple_mean3_stdErr`, which reflects the mean error from a rolling forecasting, is also important to the trend component selection.
- **Seasonality component.** The seasonality-related features `seasonal_strength`, `seas_pacf`, and `seas_acf1` are identified as important features. The feature `lumpiness` has a significant impact on seasonality component selection because it calculates the variance of variances on different blocks with the length of a seasonal period.
- **Summary.** The most important features are either from **M4metalearning** (Montero-Manso et al., 2020) or **Rcatch22** (Henderson, 2021), which justifies the suitability of using two sets of features for the tasks of component selection.

Table 4
Classification performance of our proposed method against the information criteria method with regard to the accuracy and macro F_1-score. Entries in **bold** highlight that our approach outperforms the information criteria approach.

		Accuracy	Macro F_1-score
Error component	Information criteria	85.05%	87.29%
	Feature-based	**88.35%**	**90.58%**
Trend component	Information criteria	69.73%	68.54%
	Feature-based	**74.41%**	**74.32%**
Seasonality component	Information criteria	92.27%	88.77%
	Feature-based	**95.48%**	**94.08%**
ETS model	Information criteria	55.50%	54.32%
	Feature-based	**63.51%**	**65.71%**

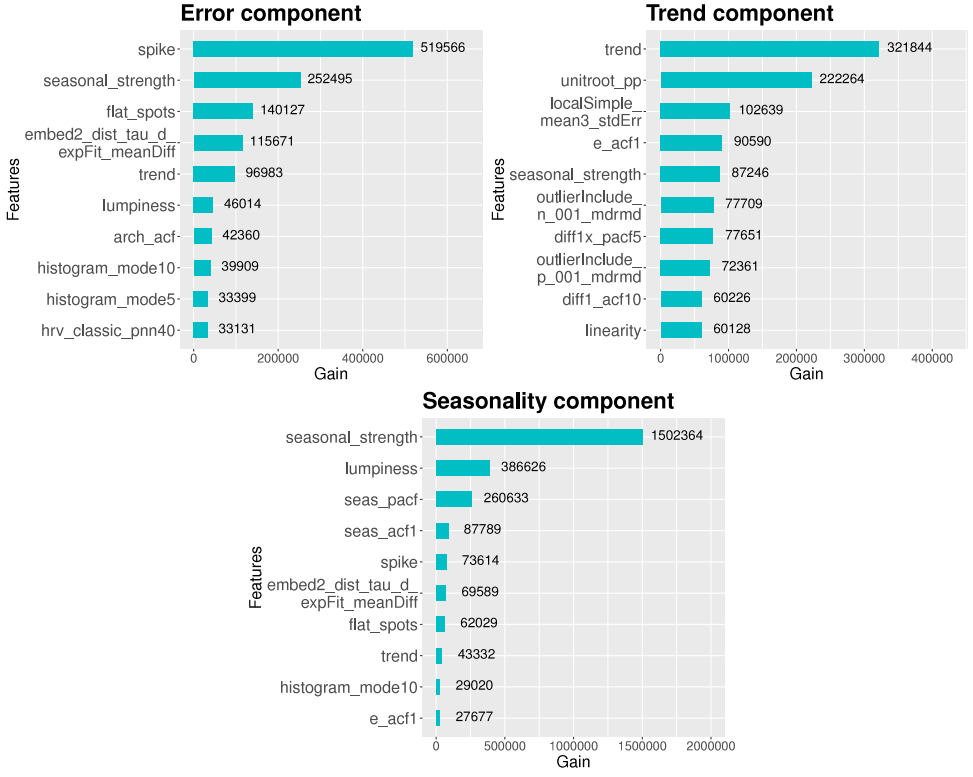

Fig. 5. Feature importance for each component selection.

4.2. Forecasting with M4 competition data

To confirm the effectiveness of our proposed approach in real data, we compare our forecasts with those from the information criteria method over the M4 dataset. Considering that ETS models are not feasible for data with multiple seasonalities (e.g., daily and hourly data) or frequencies larger than 24 (e.g., weekly data), we consider yearly, quarterly, and monthly subsets of the M4 dataset for this forecasting evaluation. The series in the M4 competition come from different fields, such as industry, finance, and economics. The M4 dataset is in the **M4comp2018** R package (Montero-Manso, Netto, & Talagala, 2018). The forecasting horizons for yearly, quarterly, and monthly data are 6, 8, and 18, respectively. The total series used in this forecasting evaluation are 23,000, 24,000, and 48,000 for yearly, quarterly, and monthly data, respectively.

We employ two widely used metrics for the point forecast evaluation: the mean absolute scaled error (MASE) (Hyndman & Koehler, 2006), and the symmetric mean absolute percentage error (sMAPE). These two metrics are calculated as

$$\text{MASE} = \frac{1}{h}\frac{\sum_{t=1}^{h} \mid y_{T+t} - \hat{y}_{T+t} \mid}{\frac{1}{T-s}\sum_{t=s+1}^{T} \mid y_t - y_{t-s} \mid},$$

$$\text{sMAPE} = \frac{1}{h}\sum_{t=1}^{h} \frac{2 \mid y_{T+t} - \hat{y}_{T+t} \mid}{\mid y_{T+t} \mid + \mid \hat{y}_{T+t} \mid},$$

$$\text{MSIS} = \frac{1}{h}\frac{\sum_{t=1}^{h}\left(y_{T+t}^{u}-y_{T+t}^{l}\right)+\frac{2}{\alpha}\left(y_{T+t}^{l}-y_{T+t}\right)\mathbb{1}\left\{y_{T+t}<y_{T+t}^{l}\right\}+\frac{2}{\alpha}\left(y_{T+t}-y_{T+t}^{u}\right)\mathbb{1}\left\{y_{T+t}>y_{T+t}^{u}\right\}}{\frac{1}{T-s}\sum_{t=s+1}^{T}|y_t-y_{t-s}|}$$

Box I.

Table 5
Performance of our proposed method against the information criteria method with regard to the mean of MASE, sMAPE, and MSIS values over M4 yearly, quarterly, and monthly data. Entries in **bold** highlight that our approach outperforms the information criteria approach.

		MASE				sMAPE				MSIS			
		1–2	3–4	5–6	1–6	1–2	3–4	5–6	1–6	1–2	3–4	5–6	1-6
Yearly	Information criteria	1.934	3.456	4.943	3.444	9.809	15.743	20.516	15.356	18.983	34.772	50.934	34.897
	Feature-based	**1.931**	**3.392**	**4.820**	**3.381**	9.933	15.746	**20.368**	**15.349**	21.956	36.878	53.386	37.407
		1–3	4–6	7–8	1–8	1–3	4–6	7–8	1–8	1–3	4–6	7–8	1-8
Quarterly	Information criteria	0.774	1.256	1.598	1.161	7.454	11.023	13.448	10.291	6.113	10.188	13.356	9.452
	Feature-based	**0.772**	**1.256**	1.599	**1.160**	**7.402**	**10.943**	**13.410**	**10.232**	6.390	10.694	14.045	9.918
		1–6	7–12	13–18	1–18	1–6	7–12	13–18	1–18	1–6	7–12	13–18	1-18
Monthly	Information criteria	0.650	0.967	1.227	0.948	9.933	13.637	17.005	13.525	5.152	8.552	11.188	8.297
	Feature-based	**0.639**	**0.959**	**1.207**	**0.935**	**9.687**	**13.385**	**16.466**	**13.179**	5.352	8.773	11.491	8.539

where $\hat{y}_t$ is the point forecast of the real value y_t of the time series at point t, h is the forecasting horizon, and s is the seasonal period of the data.

The mean scaled interval score (MSIS) (Gneiting & Raftery, 2007) as equation given in Box I, is used to measure the accuracy of a $(1-\alpha)\times 100\%$ prediction interval $\left(y_{T+t}^{l}, y_{T+t}^{u}\right)$, where $\mathbb{1}\{\cdot\}$ is an indicator function that returns a value of one if the condition is true and zero otherwise, and we set $\alpha = 0.05$.

4.3. Forecasting results

The ETS method using information criteria for model selection is implemented using the function `forecast::ets()` (Hyndman et al., 2020), with AICc as a default model selection criterion. For our method, we use the pre-trained classifiers from the simulation study for ETS model selection, and then make predictions using these selected models.

Table 5 presents the mean forecasting results. We observe the following.

- **Point forecasts.** Our method results in better point forecasts at almost all forecasting horizons for yearly (except for h = 1-4, regarding sMAPE), quarterly (except for h = 7-8, regarding MASE), and monthly data.
- **Prediction intervals.** The performance of our approach is not better than that of the information criteria method regarding the mean of MSIS.

We further provide the corresponding median values in Table 6. We can observe that our proposed method is superior to the information criteria method at almost all the forecasting horizons for yearly (except for h = 1-2, regarding MSIS), quarterly (except for h = 1-3, regarding sMAPE, and h = 7-8, regarding point forecasts), and monthly (except for h = 1-6, regarding sMAPE) data.

4.4. Forecasting accuracy analysis

We use violin plots to visualize the distribution of MASE, sMAPE, and MSIS values at all horizons, as shown in Fig. 6. We can observe that for yearly and quarterly data, the range of the MASE of our method is smaller than that of the information criteria method. For monthly data, the distributions of the forecasting accuracy of both methods are similar to each other.

Together with Tables 5 and 6, we can conclude that the proposed method resulted in some unusual forecasts that have an effect on the mean of MSIS values, as shown in the first and second plots in the third column of Fig. 6. We find that these outliers come from some unusual series that let our method select some models with multiplicative errors for them, leading to large variances of prediction errors, and thus poor prediction intervals. Some of these unusual series are available at https://github.com/Richard759/fETSmcs/blob/master/bad_case.pdf.

4.5. Significance tests

We perform Diebold–Mariano (DM) tests (Harvey, Leybourne, & Newbold, 1997) to verify whether forecasts from the proposed approach are significantly better or worse than those of the information criteria method. The null hypothesis of this test is that the two approaches have the same forecast accuracy. The alternative hypothesis is that our method is less or more accurate than the benchmark method. In these tests, if the DM tests statistic falls in the lower or upper 2.5% tail of the standard normal distribution, then we reject the null hypothesis and accept the alternative hypothesis. The DM tests are implemented using `forecast::dm.test()` (Hyndman et al., 2020) in R.

From Table 7, we can observe the following.

Table 6
Performance of our proposed method against the information criteria method with regard to the median of MASE, sMAPE, and MSIS values over M4 yearly, quarterly, and monthly data.

		MASE				sMAPE				MSIS			
		1–2	3–4	5–6	1–6	1–2	3–4	5–6	1–6	1–2	3–4	5–6	1-6
Yearly	Information criteria	1.339	2.255	3.162	2.329	5.342	8.669	11.629	8.966	7.777	13.905	20.623	15.487
	Feature-based	**1.318**	**2.214**	**3.102**	**2.289**	**5.263**	**8.426**	**11.320**	**8.721**	8.238	**13.162**	**18.278**	**15.280**
		1–3	4–6	7–8	1–8	1–3	4–6	7–8	1–8	1–3	4–6	7–8	1-8
Quarterly	Information criteria	0.573	0.910	1.123	0.886	3.513	5.598	6.835	5.608	3.993	6.199	7.658	5.977
	Feature-based	**0.570**	**0.904**	1.124	**0.886**	3.530	**5.597**	6.882	**5.600**	**3.980**	**6.115**	**7.575**	**5.939**
		1–6	7–12	13–18	1–18	1–6	7–12	13–18	1–18	1–6	7–12	13–18	1-18
Monthly	Information criteria	0.497	0.708	0.877	0.736	4.223	6.654	8.329	6.995	3.225	4.933	6.339	5.040
	Feature-based	**0.491**	**0.705**	**0.867**	**0.728**	4.247	**6.641**	**8.258**	**6.964**	**3.219**	**4.917**	**6.260**	**5.030**

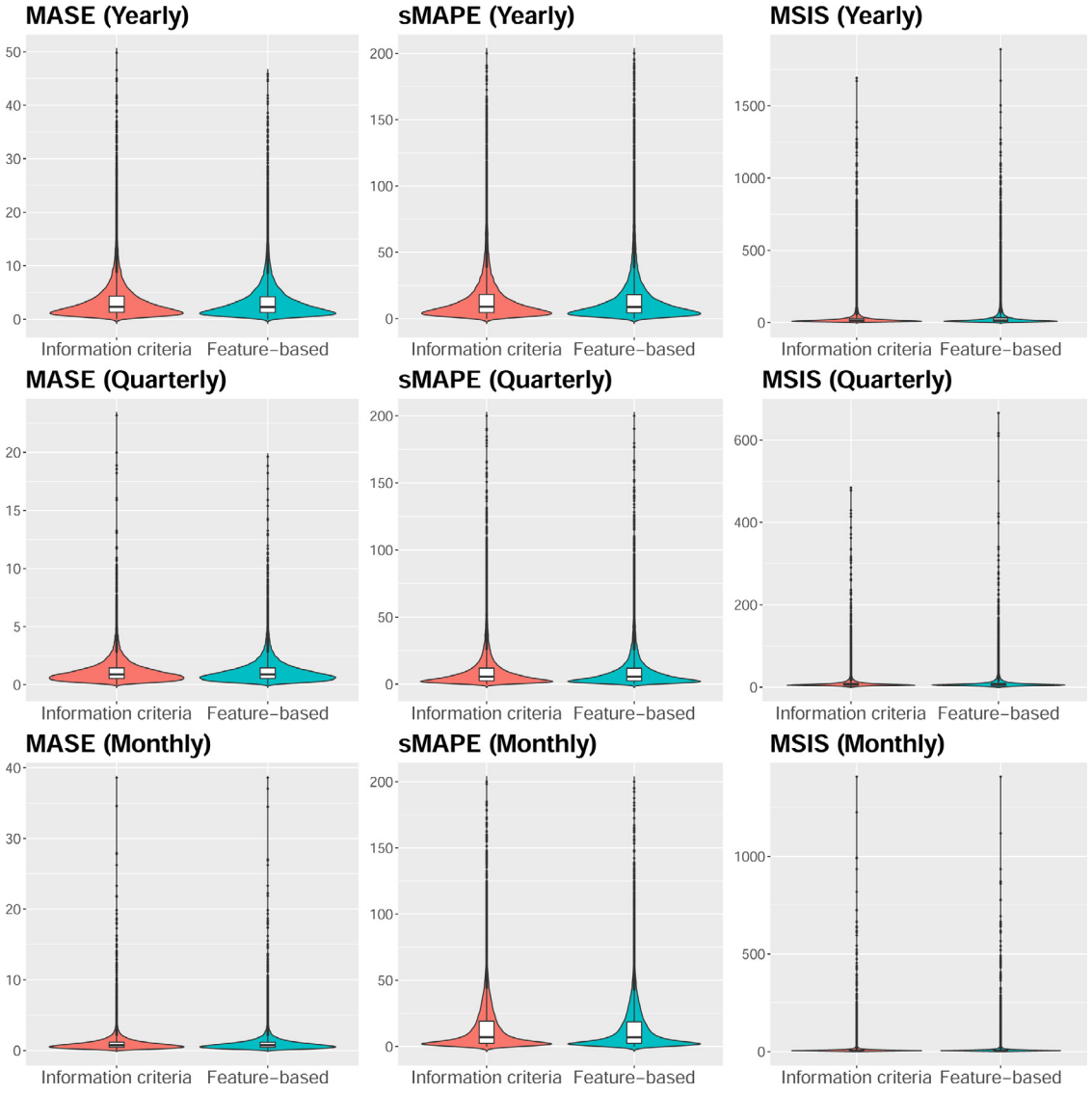

Fig. 6. Violin plots of MASE, sMAPE, and MSIS values over yearly, quarterly, and monthly data.

Table 7
Diebold–Mariano (DM) tests for comparing the predictive accuracy of the proposed method with the information criteria method. The entries show the percentage of times the forecasts of our method are significantly better or worse than the forecasts of the information criteria method at different horizons. Entries in **bold** highlight that our method is significantly better than the information criteria method.

	Yearly				Quarterly				Monthly			
	1–2	3–4	5–6	1–6	1–3	4–6	7–8	1–8	1–6	7–12	13–18	1-18
better	1.600	3.574	**5.770**	**16.839**	2.350	**9.163**	**6.329**	**17.879**	11.135	16.683	**20.410**	**22.846**
worse	1.630	3.613	5.448	15.539	2.646	8.921	6.096	17.242	11.279	17.031	19.779	22.238

Table 8
Performance of our proposed method against the information criteria method with regard to the mean of MASE, sMAPE, and MSIS values over the hospital data. Entries in **bold** highlight that our approach outperforms the information criteria approach.

	MASE				sMAPE				MSIS			
	1–6	7–12	13–18	1–18	1–6	7–12	13–18	1–18	1–6	7–12	13–18	1-18
Information criteria	0.726	0.804	0.895	0.808	16.724	18.138	19.855	18.239	4.536	5.519	6.486	5.514
Feature-based	**0.712**	**0.780**	**0.865**	**0.786**	**16.364**	**17.778**	**19.353**	**17.832**	4.556	**5.433**	**6.422**	**5.470**

Table 9
Computational time (in seconds) for producing forecasts from our method and the information criteria method in the online phase.

	M4		Hospital	
	Information criteria	Feature-based	Information criteria	Feature-based
Feature extraction	–	588	–	5
Model selection and forecasting	5782	367	67	4.1
Total	5782	955	67	9.1

- For yearly data, our approach significantly outperforms the information criteria method at long-term (h = 5-6) and all (h = 1-6) horizons.
- For quarterly data, our approach is significantly superior to the information criteria method at mid-term (h = 4-6), long-term (h = 7-8), and all (h = 1-8) horizons.
- For monthly data, our approach is significantly better than the information criteria method at long-term (h = 13-18) and all (h = 1-18) horizons.
- The proposed method results in robust forecasts at long-term and all horizons for all data.

5. Case study: Forecasting with a monthly hospital dataset

To further demonstrate the practical value of the proposed method in specific forecasting domains, we apply our method to a monthly hospital dataset. The dataset, which consists of 767 monthly series, was analyzed by Hyndman et al. (2008). The dataset is in the **expsmooth** R package (Hyndman, 2015). We save the last 18 observations for forecasting evaluation. We employ the pre-trained classifiers from the simulation study to select an appropriate ETS model for any series in this set, and then make predictions using these selected models. From Table 8, we can find that our proposed approach resulted in better forecasting results than the information criteria method at almost all horizons (except for $h = 1-6$, regarding MSIS) in terms of point forecasts and prediction intervals.

6. Computational time for producing forecasts

We recorded and compared the computational time for producing forecasts with our method and with the information criteria method over the above M4 and hospital datasets. To ensure a fair comparison, all the experiments were carried out on an AMD Ryzen 7 4800H 2.90 GHz server with eight cores. The computation time of the proposed method in the training phase for time series simulation, feature extraction, and classifier building was 1162 s, 1683 s, and 7386 s, respectively.

From Table 9, we can observe that our proposed method is significantly faster than the information criteria method at producing forecasts in the online phase. It cannot be denied that the most computationally demanding processes of our method are time series simulation, feature extraction, and classifier training in the training phase. However, this is acceptable in real application scenarios because classifiers only need to be trained once, and they can be applied to different real data without training them each time.

7. Discussion

Considering that the existing ETS method, which uses information criteria for model selection, suffers from computational complexity when applied to large-scale time series data, we proposed an efficient approach for ETS model selection that trains three classifiers on simulated data to predict model component forms. A simulation study demonstrated the model selection ability of the

Table 10
The **M4metalearning** features (Montero-Manso et al., 2020) used for ETS model component selection.

Index	Feature	Description
1	x_acf	First autocorrelation function (ACF) of a series
2	x_acf10	Sum of the squared first ten ACFs of a series
3	diff1_acf1	First ACF of a differenced series
4	diff1_acf10	Sum of the squared first ten ACFs of a differenced series
5	diff2_acf1	First ACF of a twice-differenced series
6	diff2_acf10	Sum of squared first ten ACFs of a series
7	seas_acf1	ACF of a seasonally differenced series, 0 for non-seasonal series
8	ARCH.LM	A statistic for AR conditional heteroscedasticity
9	crossing_point	The number of times a series crosses its median
10	entropy	The spectral entropy of a series
11	flat_spots	The number of flat spots in a series
12	arch_acf	Sum of squares of first 12 ACFs of a pre-whitened series
13	garch_acf	Sum of squares of first 12 ACFs of the squared residuals from a fitted GARCH(1,1) model
14	arch_r2	R^2 value of an AR model applied to a pre-whitened series
15	garch_r2	R^2 value of a GARCH(1,1) model applied to a pre-whitened series
16	hurst	Hurst
17	lumpiness	Lumpiness
18	nonlinearity	Nonlinearity
19	x_pacf5	Sum of squared first 5 partial autocorrelation functions (PACFs) of a series
20	diff1x_pacf5	Sum of squared first 5 PACFs of a differenced series
21	diff2x_pacf5	Sum of squared first 5 PACFs of a twice-differenced series
22	seas_pacf	PACF of at first seasonal lag, 0 for non-seasonal series
23	nperiods	The number of seasonal periods in a series
24	seasonal_period	The length of the seasonal period
25	trend	Trend
26	spike	Spike
27	linearity	Linearity
28	curvature	Curvature
29	e_acf1	First ACF of the remainder series from an STL decomposition
30	e_acf10	Sum of first 10 squared ACFs of the remainder series from an STL decomposition
31	seasonal_strength	Strength of seasonality
32	peak	Peak
33	trough	Trough
34	stability	Stability
35	unitRoot_kpss	A statistic for the unit root test
36	unitRoot_pp	Another statistic for the unit root test
37	series_length	Series length

proposed method. DM tests results confirmed that our approach produces robust forecasts at long-term and all horizons for yearly, monthly, and quarterly data.

The most computationally demanding processes of the proposed method are time series simulation, feature extraction, and classifier building in the training phase.

Table 11
The **Rcatch22** features (Lubba et al., 2019) used for ETS model component selection.

Index	Feature	Description
38	histogram_mode5	Mode of z-scored distribution using 5-bin histogram
39	histogram_mode10	Mode of z-scored distribution using 10-bin histogram
40	binaryStats_mean_longStretch1	Longest period of consecutive values above mean
41	outlierInclude_p_001_mdrmd	Time intervals between extreme events above mean
42	outlierInclude_n_001_mdrmd	Time intervals between extreme events below mean
43	f1ecac	First 1/e crossing of ACF
44	firstmin_ac	First minimum of ACF
45	summaries_welch_rect_area_5_1	Total power of lowest fifth of frequencies in a power spectrum
46	summaries_welch_rect_centroid	Centroid of the Fourier power spectrum
47	localSimple_mean3_stdErr	Prediction error from a rolling forecast
48	trev_1_num	A statistic for time-reversibility
49	histogramAMI_even_2_5	Automutual information
50	autoMutualInfoStats_40_gaussian_fmmi	First minimum of an information function
51	hrv_classic_pnn40	Proportion of continuous differences above 0.04σ (Mietus, Peng, Henry, Goldsmith, & Goldberger, 2002)
52	binaryStats_diff_longStretch0	Longest length of incremental decreases
53	motifThree_quantile_hh	Entropy of two continuous letters in a symbolization
54	localSimple_mean1_tauresrat	Correlation changes after iterative differencing
55	embed2_dist_tau_d_expFit_meanDiff	An exponential fit to continuous distances in 2D space
56	fluctAnal_2_dfa_50_1_2_logi_prop_r1	Proportion of slower timescale fluctuations with DFA scaling
57	fluctAnal_2_rsrangefit_50_1_logi_prop_r1	Proportion of slower timescale fluctuations with rescaled range fits
58	transitionMatrix_3ac_sumDiagCov	Trace of covariance of a transition matrix
59	periodicityWang_th0_01	A periodicity measure (Wang, Wirth, & Wang, 2007)

Table 12
Values of hyperparameters of LightGBM on the training data.

Hyperparameters	f_e	f_t	f_s
`eta`	0.05	0.05	0.05
`num_leaves`	92	80	64
`min_data_in_leaf`	90	100	60
`max_bin`	175	225	175
`num_boost_round`	600	800	1000
`bagging_fraction`	0.8	0.8	0.8
`bagging_freq`	4	4	4
`feature_fraction`	0.7	0.7	0.7
`force_col_wise`	True	True	True
`bagging_seed`	123	123	123

However, this is acceptable in practice because classifiers for component form prediction only need to be trained once, and they can be applied to different real data without retraining. The good forecasting results on a hospital dataset from a specific domain demonstrated the robustness of the proposed method. We also showed that the proposed method is more efficient at producing forecasts over the M4 and hospital datasets in the online phase.

However, the good forecasting performance of the proposed approach depends on the selection of an appropriate set of time series features and careful parameter tuning. Compared with the information criteria method, our method achieved impressive performance, but at the cost of interpretability. That is, we have no idea about how well a selected model from our classifiers fits the data and how complicated the model is.

Our approach feature-based ETS model component selection (fETSmcs) is implemented as an `R` package. Our code is open-source and publicly available at https://github.com/Richard759/fETSmcs.

8. Concluding remarks

To tackle the inefficiency issue in current big data settings, we explored the use of features for ETS model selection by training classifiers on simulated data to predict component forms. We provided a simulation study, as well as two real data applications, to demonstrate the effectiveness of the proposed method.

However, the proposed approach depends on the manual choice of an appropriate set of features. Some research (e.g., Li et al., 2020) introduces automatic features for forecast model combination, and one potential future research direction would be to adopt automatic features for ETS model selection. Another potential avenue for future research would be to design a model selection criterion that incorporates information criteria and feature-based selection, aiming at achieving impressive performance while still maintaining good interpretability.

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

We are grateful to the editor and two anonymous reviewers for their helpful comments that improved the contents of this paper. We are also grateful to Professor Yanfei Kang from Beihang University for her feedback in the early stages of the development of the proposed method in this paper.

Appendix A. Feature description

Tables 10 and 11 introduce the 59 used features and their corresponding descriptions.

Appendix B. Experimental setup for LightGBM

We performed hyperparameter optimization with a grid search by measuring the area under the receiver operating characteristic curve (AUC) on a 20% holdout version of the training data in a five-fold cross-validation procedure. The searching ranges of four hyperparameters are as follows.

- `num_leaves`: The number of leaves ranges from 8 to 128.
- `min_delta_in_leaf`: The minimum number of the records a leaf may have, ranging from 10 to 110.
- `max_bin`: The maximum number of bins that feature values will be bucketed in, ranging from 5 to 255.
- `num_boost_round`: The number of boosting stages, ranging from 100 to 1100.

Considering computation time, we set fixed values for some other hyperparameters. Table 12 reports the optimal values of four hyperparameters of LightGBM, as well as other hyperparameters, with fixed values for the three classifiers f_e, f_t, and f_s.

References

Billah, B., King, M. L., Snyder, R. D., & Koehler, A. B. (2006). Exponential smoothing model selection for forecasting. *International Journal of Forecasting*, *22*(2), 239–247.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Brown, R. G. (1956). *Exponential smoothing for predicting demand*. Little, Cambridge, Massachusetts: NBER Cambridge, MA, USA.

Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw/Hill.

Collopy, F., & Armstrong, J. S. (1992). Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. *Management Science*, *38*(10), 1394–1414.

Fildes, R., & Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, *68*(8), 1692–1701.

Fulcher, B. D., & Jones, N. S. (2017). Hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell Systems*, *5*(5), 527–531.

Gardner Jr., E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, *4*(1), 1–28.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291.

Henderson, T. (2021). Rcatch22: Calculation of 22 Canonical time-series characteristics. R package version 0.1.12.

Hyndman, R. J. (2015). Expsmooth: Data sets from "forecasting with exponential smoothing". R package version: 2.3.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., et al. (2020). Forecast: Forecasting functions for time series and linear models. URL http://pkg.robjhyndman.com/forecast, R package version 8.14.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439–454.

Kang, Y., Hyndman, R. J., & Li, F. (2020). GRATIS: GeneRAting time series with diverse and controllable characteristics. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *13*(4), 354–376.

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*(2), 345–358.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3146–3154.

Li, X., Kang, Y., & Li, F. (2020). Forecasting with time series imaging. *Expert Systems with Applications*, *160*, Article 113680.

Li, X., Petropoulos, F., & Kang, Y. (2022). Improving forecasting by subsampling seasonal time series. *International Journal of Productions Research*, 1–17.

Lubba, C. H., Sethi, S. S., Knaute, P., Schultz, S. R., Fulcher, B. D., & Jones, N. S. (2019). Catch22: Canonical time-series characteristics. *Data Mining and Knowledge Discovery*, *33*(6), 1821–1852.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74.

Meira, E., Oliveira, F. L. C., & Jeon, J. (2021). Treating and pruning: New approaches to forecasting model selection and combination using prediction intervals. *International Journal of Forecasting*, *37*(2), 547–568.

Mietus, J., Peng, C., Henry, I., Goldsmith, R., & Goldberger, A. (2002). The pNNx files: Re-examining a widely used heart rate variability measure. *Heart*, *88*(4), 378–380.
Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86–92.
Montero-Manso, P., Netto, C., & Talagala, T. S. (2018). M4comp2018: Data from the M4-competition. R package version: 0.1.0.
Nikolopoulos, K., & Petropoulos, F. (2018). Forecasting for big data: Does suboptimality matter? *Computers & Operations Research*, *98*, 322–329.
Ord, J. K., Koehler, A. B., & Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, *92*(440), 1621–1629.
Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, *38*(3), 705–871.
Petropoulos, F., & Grushka-Cockayne, Y. (2021). Fast and frugal time series forecasting. Available At SSRN 3792565.
Reid, D. (1972). A comparison of forecasting techniques on economic time series. Forecasting in Action. Operational Research Society and the Society for Long Range Planning.
Seaman, B. (2018). Considerations of a retail forecasting practitioner. *International Journal of Forecasting*, *34*(4), 822–829.
Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020). Are forecasting competitions data representative of the reality? *International Journal of Forecasting*, *36*(1), 37–53.
Svetunkov, I. (2021). Smooth: Forecasting using state space models. URL https://github.com/config-i1/smooth, R package version 3.1.1.
Talagala, T. S., Hyndman, R. J., Athanasopoulos, G., et al. (2018). Meta-learning how to forecast time series. *Monash econometrics and business statistics working papers*, *6*, (p. 18). Monash University, Department of Econometrics and Business Statistics.
Talagala, T. S., Li, F., & Kang, Y. (2021). FFORMPP: Feature-based forecast model performance prediction. *International Journal of Forecasting (in Press)*.
Wang, X., Kang, Y., Petropoulos, F., & Li, F. (2021). The uncertainty estimation of feature-based forecast combinations. *Journal of the Operational Research Society*, 1–15.
Wang, X., Wirth, A., & Wang, L. (2007). Structure-based statistical features and multivariate time series clustering. In *Seventh IEEE international conference on data mining* ICDM 2007, (pp. 351–360). IEEE.
Zhou, Z.-H. (2021). *Machine learning*. Springer.