Machine Learning (2022) 111:739–789
https://doi.org/10.1007/s10994-021-06126-z

Model selection in reconciling hierarchical time series

Mahdi Abolghasemi[1] · **Rob J. Hyndman**[2] · **Evangelos Spiliotis**[3] · **Christoph Bergmeir**[1]

Received: 16 October 2020 / Revised: 13 August 2021 / Accepted: 3 November 2021 /
Published online: 14 January 2022

Abstract

Model selection has been proven an effective strategy for improving accuracy in time series forecasting applications. However, when dealing with hierarchical time series, apart from selecting the most appropriate forecasting model, forecasters have also to select a suitable method for reconciling the base forecasts produced for each series to make sure they are coherent. Although some hierarchical forecasting methods like minimum trace are strongly supported both theoretically and empirically for reconciling the base forecasts, there are still circumstances under which they might not produce the most accurate results, being outperformed by other methods. In this paper we propose an approach for dynamically selecting the most appropriate hierarchical forecasting reconciliation method and leading to more accurate coherent forecasts. The approach, which we call conditional hierarchical forecasting, is based on machine learning classification methods that use time series features to select the reconciliation method for each hierarchy. Moreover, it allows the selection to be tailored according to the accuracy measure of preference and the hierarchical level(s) of interest. Our results suggest that conditional hierarchical forecasting can lead to significantly more accurate forecasts than standard approaches, especially at lower hierarchical levels.

Keywords Hierarchical forecasting · Machine learning · Time series features · Classification

Editor: Gustavo Batista.

✉ Mahdi Abolghasemi
mahdi.abolghasemi@monash.edu

Rob J. Hyndman
rob.hyndman@monash.edu

Evangelos Spiliotis
spiliotis@fsu.gr

Christoph Bergmeir
Christop.bergmeir@monash.edu

[1] Department of Data science & AI, Monash University, Clayton, Melbourne, Australia

[2] Department of Econometrics and Business Statistics, Monash University, Clayton, Melbourne, Australia

[3] Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece

1 Introduction and background

Forecasting is essential for supporting decision-making, especially in applications that involve a lot of uncertainty. For instance, accurately forecasting the future demand of stock keeping units (SKUs) can significantly improve supply chain management (Ghobbar and Friend, 2003), reduce inventory costs (Syntetos et al., 2010), and increase service levels (Pooya et al., 2019), particularly under the presence of promotions (Giir Ali et al., 2009; Abolghasemi et al., 2020). In order to obtain more accurate forecasts, forecasters typically try to identify the most appropriate forecasting model for each series from a variety of alternatives. Although this task can provide significant improvements under perfect foresight (Fildes, 2001), it is difficult to effectively perform in practice due to model, parameter, and data uncertainty (Petropoulos et al., 2018). Thus, many strategies have been proposed in the literature to effectively perform forecasting model selection (Fildes and Petropoulos, 2015), most of which are based on the in-sample and out-of-sample accuracy of the forecasting models (Tashman, 2000), their complexity (Hyndman et al., 2002), and the features that time series display (Montero-Manso et al., 2020; Petropoulos et al., 2014).

However, in business forecasting applications, data is typically grouped based on its context and characteristics, thus structuring cross-sectional hierarchies. For example, although the demand of an SKU can be reported at a store level, it can be also aggregated (summed) at a regional or national level. Similarly, demand can be aggregated for various SKUs of the same type (e.g., dairy products) or category (e.g., foods). As a result, hierarchical time series introduce additional complexity to the whole forecasting process since, apart from selecting the most appropriate forecasting model for each series, forecasters have also to account for coherence, i.e. make sure that the forecasts produced at the lower hierarchical levels will sum up to those produced at the higher ones (Athanasopoulos et al., 2020). In fact, coherence is a prerequisite in hierarchical forecasting (HF) applications as it ensures that different decisions made across different hierarchical levels will be aligned.

Naturally, the demand recorded at lower hierarchical levels will always add up to the observed demand at higher levels. However, this is rarely the case for forecasts which are usually produced for each series separately and are therefore incoherent. To achieve coherence, various HF methods can be used for reconciling the individual, base forecasts (Spiliotis et al., 2019). The most basic HF method is probably the bottom-up (BU), according to which base forecasts are produced just for the series at the lowest level of the hierarchy, and are then aggregated to provide forecasts for the series at the higher levels (Dangerfield and Morris , 1992). Top-down (TD) is another option which involves forecasting just the series at the highest level of the hierarchy and then using proportions to disaggregate these forecasts and predict the series at the lower levels Gross and Sohl 1990; Athanasopoulos et al. 2009. Middle-out (MO) mixes the above-mentioned methods, producing base forecasts for a middle level of the hierarchy and then aggregating or disaggregating them to forecast the higher and lower levels, respectively (Abolghasemi et al. , 2019). Finally, a variety of HF methods that combine (COM) the forecasts produced at all hierarchical levels have been proposed in the literature, usually resulting in coherent and more accurate forecasts (Hyndman et al., 2011; Wickramasuriya et al., 2019; Jeon et al. , 2019).

From the HF methods found in the literature, a COM method, called minimum trace (Wickramasuriya, 2019, MinT;][), has been distinguished due to the strong theory supporting it and the results of many empirical studies highlighting its merits over other alternatives (Abolghasemi et al., 2019; Burba and Chen, 2021; Spiliotis et al., 2020). However, there are still circumstances under which MinT might fail to provide the most accurate forecasts. For instance, since MinT is based on the estimation of the one-step-ahead error covariance matrix, the method might be proven inappropriate when the in-sample errors of the baseline forecasting models do not represent post-sample accuracy, the assumption that the multi-step forecast error covariance is proportional to the one-step forecast error covariance is unrealistic, or the required estimations are computationally too hard to make. Moreover, since MinT treats all levels equally, it is not optimized with respect to certain hierarchical levels of interest. Finally, given that medians are not additive, there is no reason to expect that MinT will always improve the mean absolute forecast error, or other accuracy measures that are based on absolute forecast errors.

In such cases, simpler HF methods like the BU and the TD may be useful. However, there is inadequate evidence about which of the two methods to use (Hyndman et al., 2011). For example, the BU method is typically regarded as more suitable for short-term forecasts and for hierarchies in which bottom series are not highly correlated and not dominated by noise (Kahn, 1998). On the other hand, the TD method is usually regarded as more appropriate for long-term forecasts, but less accurate for predicting the series at the lower aggregation levels due to information loss (Dangerfield and Morris, 1992; Kahn, 1998). It seems that no reconciliation method can fit all kinds of HF problems and that, similarly to forecasting model selection, the appropriateness of the different HF methods depends on various factors, including the particularities of the time series (Nenova and May, 2016) and the structure of the hierarchy (Abolghasemi et al., 2019; Fliedner, 1999; Fliedner , 2001; Gross and Sohl, 1990). The above findings reconfirm highlight the potential benefits of conditional hierarchical forecasting (CHF); i.e. the improvements in terms of forecasting accuracy that could be possibly achieved if forecasters were able to select the most appropriate HF method according to the characteristics of the series that form a hierarchy. In this paper we propose an approach for performing such a conditional selection using time series features as leading indicators (Kang et al., 2017; Spiliotis et al., 2020a) and machine learning (ML) methods for conducting the classification. Essentially, we suggest that the forecasting accuracy of the different HF methods found in the literature is closely related with the characteristics of the individual series and that, based on these relationships, “horses for courses” can be effectively identified (Petropoulos et al., 2014). In addition, CHF allows the selection to be tailored according to the accuracy measure of preference (e.g. mean absolute or squared error) and the hierarchical level(s) of interest (e.g. top or bottom level), thus adapting to the requirements of the examined forecasting task and effectively supporting decisions.

Table 1 summarizes major studies conducted in the field of HF, putting a particular emphasis on approaches proposed in the literature for reconciling base forecasts using either combination or selection methods. The method proposed in this paper is also included in the table to facilitate comparisons. As seen, various studies have considered ML methods for performing TD (Mancuso et al., 2020), MO (Abolghasemi et al., 2019), or BU (Burba and Chen, 2021; Spiliotis et al., 2020) reconciliation in a dynamic fashion by using base forecasts and explanatory variables as input to regression models, including neural networks (NN), regression trees (RT), and support vector machines (SVM), among others. Instead of reconciling the base forecasts directly, other studies have proposed combining the reconciled forecasts produced by standard HF methods using simple weighting schemes (Abouarghoub et al., 2018) or selecting the most appropriate one from a list of alternatives (Nenova and May, 2016). As such, the spirit of our work is similar to that of Nenova and May (2016) since both studies aim to select the most suitable reconciliation method for a hierarchy of interest. However, we take a different approach in doing so. First, in contrast to (Nenova and May, 2016), which exploits time series correlations and rank

Table 1 State-of-the-art: Major studies conducted in the field of hierarchical forecasting for reconciling base forecasts using either combination or selection approaches

Evaluation	Approach	HF Method	Data	Frequency	#levels	# hierarchies	Features	Refs
Theoretical & Empirical	Linear combination of base forecasts	COM	Tourism demand	Quarterly, Monthly	4	1, 1	Base forecasts	(Hyndman et al., 2011; Wickramasuriya et al., 2019)
Empirical	Combination of HF methods (BU, MO & TD)	Equal or scaled weights	Shipping freight earnings	Monthly	4	1	Reconciled forecasts	(Abouarghoub et al., 2018)
Empirical	Selection of HF method (BU or TD)	ML classification (LR, LDA, RT, SVM & C5)	Product sales	Monthly	2	56	Correlation & rank predictors	(Nenova and May, 2016)
Empirical	Dynamic MO	ML regression (XGB, MLP & SVM)	Foods sales	Weekly	3	61	Base forecasts & explanatory variables	(Abolghasemi et al., 2019)
Empirical	Dynamic TD	ML regression (MLP & CNN)	Supermarket sales, Electricity demand	Daily, 10 minutely	3, 2	1,1,1	Time series data & explanatory variables	(Mancuso et al., 2020)
Empirical	Dynamic BU	ML regression (RF & XGB)	Tourism demand, Foods sales	Quarterly, Weekly	4, 3	1, 55	Base forecasts	(Spiliotis et al., 2020)
Empirical	Dynamic BU	ML regression (encoder-decoder NN)	Dairy sales, Walmart sales, Tourism demand, Electricity demand	Weekly, Daily, Quarterly, Yearly	2, 3, 4, 4	1, 1, 1, 1	Base forecasts	(Burba and Chen, 2021)
Empirical	Selection of HF method (BU, TD, or COM)	ML classification (RF, SVM & XGB)	Foods sales, Prison population, Tourism demand	Weekly, Quarterly, Quarterly	3, 3, 3	55, 1, 1	Time series features	This paper

predictors (i.e., features related to the structure of the hierarchy), our selection is performed using a comprehensive set of time series features that describe the behavior of the individual series comprising the hierarchy. Time series features have been minimally considered in some other studies that evaluated the impact of the series autocorrelation (Chen and Boylan, 2009), demand type (Widiarta et al., 2007; Widiarta et al., 2008), and forecasting horizon (Burba and Chen, 2021) on the appropriateness of the BU and the TD methods, mostly using simulations and ex-post evaluations. Second, in our study we use a different set of baseline HF methods, including COM in addition to the TD and BU methods. This is done because several studies have shown that COM can outperform standard HF methods, being also significantly different in nature than BU and TD in terms of the approach used for performing the reconciliation (Hyndman et al., 2016; Hyndman et al., 2011; Abolghasemi et al., 2020). Third, we use a different set of models for conducting the classification, including more advanced decision-tree-based algorithms, such as random forests (RF) and eXtreme Gradient Boosting (XGB), that have shown promising results in various forecasting tasks and competitions (Montero-Manso et al., 2020; Chen and Guestrin, 2016). Fourth, we evaluate the performance of our method by considering diverse sets of hierarchical data and optimization criteria in terms of the hierarchical level at which the forecasts should be considered as optimal, the characteristics of time series, the measure used for assessing accuracy, and the forecasting horizon. We also conduct an empirical comparison between the method proposed in this paper and the one described in Nenova and May (2016), and show the importance of time series characteristics in selecting the most appropriate reconciliation method.

We benchmark the accuracy of the proposed approach against various HF methods, both standard and state-of-the-art, considering a variety of optimization criteria, and using three large data sets from the retail, tourism and justice sectors. Our results suggest that CHF leads to superior forecasts that outperform those of the individual HF methods examined. Thus, we conclude that selection should not be limited to forecasting models, but be expanded to HF methods as well.

The remainder of the paper is organized as follows. Section 2 describes the most popular HF methods found in the literature and Sect. 3 introduces CHF. Section 4 presents the primary data set used for the empirical evaluation of the proposed approach and describes the experimental set-up. Section 5 presents the results of the experiment and discusses our findings. Finally, Sect. 6 concludes the paper.

2 Hierarchical forecasting methods

In this section, we discuss the TD, BU, and COM as three well-established HF methods that are widely used in the literature and in practice for reconciling hierarchical base forecasts. These methods are also the ones considered in this study, both as alternatives of the conditional HF approach to be described in the next section and as benchmarks. For a more detailed discussion on the existing HF methods, their advantages, and drawbacks, please refer to the study of Athanasopoulos et al. (2020).

Before proceeding, we introduce the following notations and parameters that will facilitate the discussion of the three methods:

m:	Total number of series in the hierarchy
m_i:	Total number of the series for level i;

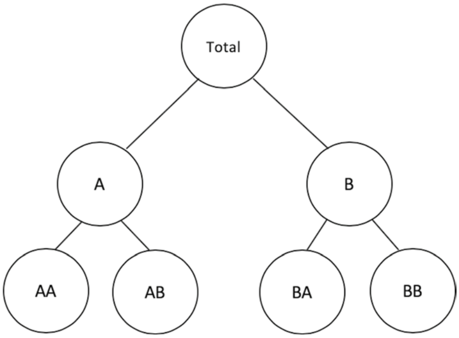

Fig. 1 A three-level hierarchical structure

k: Total number of the levels in hierarchy;
n: Number of the observations in each series;
$Y_{x,t}$: The t^{th} observation of series Y_x;
$\hat{Y}_{x,n}(h)$: h—step-ahead independent base forecast of series Y_x based on n observations;
$\boldsymbol{Y}_{i,t}$: The vector of all observations at level i;
$\hat{\boldsymbol{Y}}_{i,t}(h)$: h—step-ahead forecast at level i;
$\boldsymbol{Y}_t$: A column vector including all observations;
$\hat{\boldsymbol{Y}}_n(h)$: h—step-ahead independent base forecast of all series based on$n$ observations;
$\tilde{\boldsymbol{Y}}_n(h)$: The final reconciled forecasts of all series

We can express a hierarchical time series as $\boldsymbol{Y}_t = \boldsymbol{S}\boldsymbol{Y}_{k,t}$, where $\boldsymbol{S}$ is a summing matrix of order $m \times m_k$. For example, we can express the three-level hierarchical time series shown in Fig. 1 as:

$$\begin{bmatrix} Y_t \\ Y_{A,t} \\ Y_{B,t} \\ Y_{AA,t} \\ Y_{AB,t} \\ Y_{BA,t} \\ Y_{BB,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ & & \boldsymbol{I}_4 & \end{bmatrix} \times \begin{bmatrix} Y_{AA,t} \\ Y_{AB,t} \\ Y_{BA,t} \\ Y_{BB,t} \end{bmatrix}$$

Accordingly, we can express various hierarchical structures with a unified format as $\tilde{\boldsymbol{Y}}_n(h) = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{Y}}_n(h)$, where $\boldsymbol{G}$ is a matrix of order $m \times m_k$ which elements depend on the type of the reconciliation method used, in our case the BU, TD, and COM methods Hyndman and Athanasopoulos (2021).

2.1 Bottom-up

BU is the simplest HF method according to which we forecast the series at the bottom level of the hierarchy and then aggregate these forecasts to obtain forecasts at higher levels. In this case, the matrix $\boldsymbol{G}$ can be constructed as $\boldsymbol{G} = [\mathbf{0}_{m_k \times (m-m_k)} | \boldsymbol{I}_{m_k}]'$, where $\mathbf{0}_{i \times j}$ is a $i \times j$ null matrix.

2.2 Top-down

In the TD method, base forecasts are produced at the top level of the hierarchy and then disaggregated to the lower levels with appropriate factors. While there are various ways for computing such factors and disaggregating the top level forecasts, we consider the proportions of the historical averages since it is a widely used alternative that provides reasonable results (Athanasopoulos et al., 2009). These proportions are computed as follows

$$p_j = \frac{\sum_{t=1}^{n} Y_{j,t}}{\sum_{t=1}^{n} Y_t}, \qquad j = 1, \dots, m_k \tag{1}$$

where p_j represents the average of the historical value of the bottom level series $Y_{j,t}$ relative to the average value of the total aggregate Y_t. We can then construct the vector $\boldsymbol{g} = [p_1, p_2, p_3, \dots, p_{m_k}]$ and matrix $\boldsymbol{G} = [\boldsymbol{g} \mid \mathbf{0}_{m_k \times (m-1)}]'$.

2.3 Optimal combination

The COM method produces base forecasts for all series across all hierarchical levels and then combines them with a linear model to obtain the reconciled forecasts. Suppose

$$\tilde{\boldsymbol{Y}}_n(h) = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{Y}}_n(h)$$

depicts the h-step-ahead reconciled forecasts. Then, the covariance matrix of the errors of these forecasts can be given by

$$\boldsymbol{V}_h = \mathrm{Var}[\boldsymbol{y}_{n+h} - \tilde{\boldsymbol{Y}}_n(h)] = \boldsymbol{S}\boldsymbol{G}\boldsymbol{W}_h\boldsymbol{G}'\boldsymbol{S}',$$

where $\boldsymbol{W}_h$ is the variance-covariance matrix of the h-step-ahead base forecast errors (Wickramasuriya et al., 2019; Hyndman et al., 2016). It can be shown that the matrix $\boldsymbol{G}$ that minimizes the trace of $\boldsymbol{V}_h$ such that it generates unbiased reconciled forecasts, i.e., $\boldsymbol{SGS} = \boldsymbol{S}$, is given by

$$\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}_h^{\dagger}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{\dagger},$$

where $\boldsymbol{W}_h^{\dagger}$ is the generalized inverse of $\boldsymbol{W}_h$.

There are a few different ways to estimate $\boldsymbol{W}_h$. In this study we consider the shrinkage estimation as it has been empirically shown that it yields the most accurate forecasts in many HF applications (Abolghasemi et al., 2019; Spiliotis et al., 2020; Wickramasuriya et al., 2019). Using the shrinkage method, this matrix can be estimated by $\boldsymbol{W}_h = k_h\left(\lambda_D\hat{\boldsymbol{W}}_{1,D} + (1-\lambda_D)\hat{\boldsymbol{W}}_1\right)$. The diagonal target of the shrinkage estimator is $\hat{\boldsymbol{W}}_{1,D} = \mathrm{diag}(\hat{\boldsymbol{W}}_1)$ and the shrinkage parameter is given by

$$\lambda_D = \frac{\sum_{i \neq j} \mathrm{Var}(\hat{r}_{ij})}{\sum_{i \neq j} \hat{r}_{ij}^2},$$

where $\hat{r}_{ij}$ is the $(i,j)^{th}$ element of the one-step-ahead in-sample correlation matrix (Schafer and Strimmer 2005).

The COM method was implemented using the `MinT` function of the *hts* package for Hyndman et al. (2020b).

3 Conditional hierarchical forecasting

Time series often depict different patterns, such as seasonality, randomness, noise, and auto-correlation (Kang et al., 2017). As a result, there is no model that can consistently forecast all types of series more accurately than other models, even relatively simple ones (Petropoulos et al., 2014; Fildes and Petropoulos, 2015). Similarly, although some models may perform better on a time series data set of particular characteristics, there is no guarantee that this will always be the case (Spiliotis et al., 2020a). For example, although exponential smoothing (Gardner, 1985) typically produces relatively accurate forecasts for seasonal series, it might be outperformed by ML methods when a large number of observations is available (Smyl , 2020). Thus, selecting the most appropriate forecasting model for each series becomes a challenging task for improving overall forecasting accuracy (Montero-Manso et al., 2020).

Model selection has been extensively studied in the forecasting literature. Although there is no unique way to determine the most appropriate forecasting model for each series, empirical studies have provided effective strategies for performing this task (Fildes and Petropoulos, 2015). From these strategies, the approaches that build on time series features are among the most promising given that the latter can effectively represent the behaviour of the series in an abstract form and match it with the relative performance of various forecasting models (Reid, 1972; Meade, 2000; Wang et al., 2009; Petropoulos et al., 2014; Kang et al., 2017; Abolghasemi et al., 2020).

Expert systems and rule-based forecasting were two of the early approaches to be suggested for forecasting model selection (Collopy and Armstrong, 1992; Mahajan and Wind, 1988). Collopy and Armstrong (1992) considered domain knowledge along with 18 time series features and proposed a framework that consisted of 99 rules to select the most appropriate forecasting model from 4 alternatives. In another study, Adya et al. (2000) considered 6 features and 64 rules to select the most accurate forecasting model from 3 alternatives. Similarly, Adya et al. (2001) proposed an approach to automatically extract time series features and choose the best forecasting model. Petropoulos et al. (2014) measured the impact of 7 time series features plus the length of the forecasting horizon on the accuracy of 14 popular forecasting models, while (Kang et al., 2017) and (Spiliotis et al., 2020a) linked the performance of standard time series forecasting models with that of various indicative features using data from well-known forecasting competitions. More recently, Montero-Manso et al. (2020) used 42 time series features to determine the weights for optimally combining 9 different forecasting models, winning the second place in the M4 forecasting competition (Makridakis et al., 2020).

Inspired by the work done in the area of forecasting model selection, we posit that HF methods can be similarly selected using time series features and, based on such a selection, improve forecasting accuracy for the case of hierarchical series, while simultaneously reducing computational cost (for more details on this topic please refer to Appendix E). In this respect, we proceed by computing various time series features across all hierarchical levels and propose using these features for selecting the HF method that best suites the hierarchy, i.e., produces on average the most accurate forecasts for all the series it

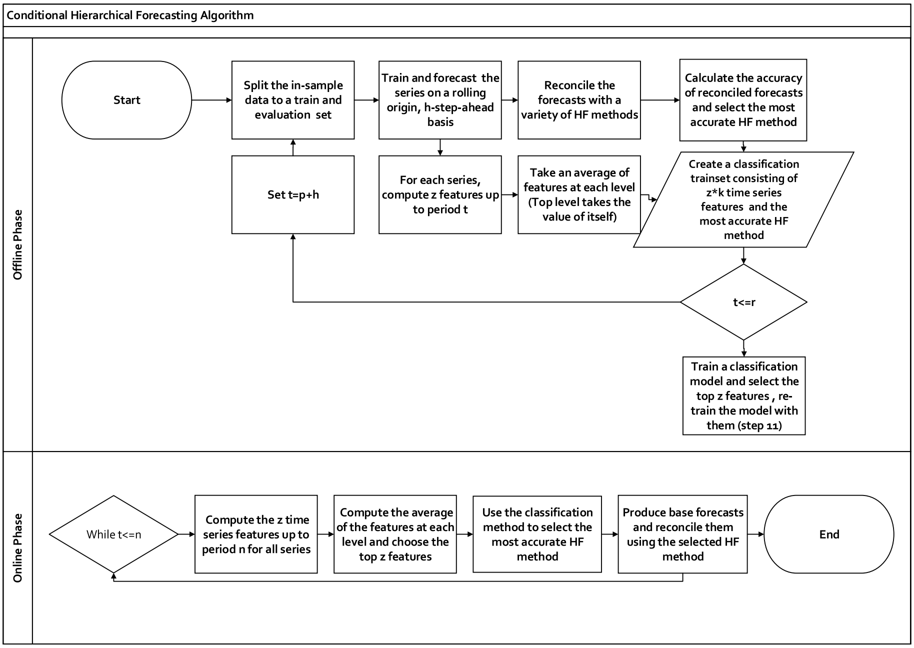
Conditional Hierarchical Forecasting Algorithm
Offline Phase
Start
Split the in-sample data to a train and evaluation set
Train and forecast the series on a rolling origin, h-step-ahead basis
Reconcile the forecasts with a variety of HF methods
Calculate the accuracy of reconciled forecasts and select the most accurate HF method
Set t=p+h
For each series, compute z features up to period t
Take an average of features at each level (Top level takes the value of itself)
Create a classification trainset consisting of z*k time series features and the most accurate HF method
t<=r
Train a classification model and select the top z features, re-train the model with them (step 11)
Online Phase
While t<=n
Compute the z time series features up to period n for all series
Compute the average of the features at each level and choose the top z features
Use the classification method to select the most accurate HF method
Produce base forecasts and reconcile them using the selected HF method
End

Fig. 2 CHF algorithm flowchart

comprises (Abolghasemi et al., 2020; Petropoulos et al., 2014)[1]. The selection is done by employing a popular ML classification method. The proposed approach, to be called conditional HF (CHF), is summarised in Appendix A and presented in the flowchart of Fig. 2.

Given that CHF builds on time series features and its accuracy is directly connected with the representatives of the features used, as well as the capacity of the algorithm employed for selecting the most appropriate HF method, it becomes evident that choosing a set of diverse, yet finite features is a prerequisite for enhancing the performance of the proposed classification approach. There are many features that can be used to describe time series patterns. For example, Fulcher et al. (2013) extracted more than 7,700 features for describing the behavior of the time series and then summarized them into 22 canonical features, losing just 7% of accuracy in a classification task (Lubba et al. 2019). Similarly, Wang et al. (2006) and Kang et al. (2017) suggested that a relatively small number of features can be used for effectively visualizing time series and performing forecasting model selection. Based on the above, we decided to consider 32 features for the CHF method so that the patterns of the hierarchical series are effectively captured without exaggerating. These features, described in Appendix B, included *entropy, lumpiness, stability, hurst, seasonal-period, seasonal-strength, trend, curvature, e-acf1, e-acf10, x-acf1, x-acf10, diff1-acf1, diff1-acf10, diff2-acf1, diff2-acf10, seas-acf1, x-pacf5, diff1x-pacf5, diff2x-pacf5, seas-pacf, linearity, non-linearity, max-var-shift, max-kl-shift, fluctanal-prop-r1, unitroot-kpss, arch-acf, garch-acf, arch-r2, garch-r2, and arch-test*, and were computed using the *tsfeatures* package for Hyndman

[1] Forecasting accuracy is first measured for each hierarchical level separately. Then, forecast errors are averaged again to measure the accuracy across the complete hierarchy.

et al. (2019). Our list of features is mostly inspired by recent studies that have successfully used time series features to develop meta-learning forecasting algorithms, e.g. for model selection and combination, (Montero-Manso et al., 2020), being tailored however for the particular requirements of the conditional hierarchical forecasting task. For more details about these features, please refer to the studies of Wang et al. (2006) and Kang et al. (2017).

Note that CHF is flexible in terms of the method that will be employed for performing the classification. That is, users can select their classification method of choice for identifying the most accurate HF method and reconciling the base forecasts produced for the examined hierarchy. In this study we considered seven methods, namely logistic regression (LR), linear discriminant analysis (LDA), regression tree (RT), RF, XGB, SVM, and neural network (NN). However, we decided to include only XGB in the main part of the paper since, in most cases, the other classification models performed worse than XGB, a powerful classification method that has been successfully applied in various forecasting and classification problems (Chatzis et al., 2018; Demolli et al., 2019; Nielsen , 2016). Appendix C summarizes the results of our experiments when classifiers other than XGB are used for implementing CHF.

Observe that CHF is also flexible in terms of the forecasting models that will be used for producing the base forecasts and the HF methods that will be considered for their reconciliation. For the latter case, we considered three HF methods (BU, TD, and COM), as described in Sect. 2. The reasoning is two-fold. First, classification methods tend to perform better when the number of classes is limited and, as a result, the key differences between the classes are easier to identify (Hastie and Tibshirani 1998). Second, we believe that the selected HF methods are diverse enough, each one focusing on different levels of the hierarchy and adopting a significantly different approach for reconciling the base forecasts. Although we could have considered more HF methods of those proposed in the literature, they are mostly variants of the examined three methods (especially the COM method) and are therefore sufficiently covered.

We should also highlight that the rolling origin evaluation of the off-line phase can be adjusted to any desirable set-up that might be more suitable to the user. For example, if computational cost is not an issue, instead of updating the forecast origin by h periods at a time, a step of one period could be considered to further increase the size of the set used for training the classification method and facilitate learning. The main motivation for considering an h-step-ahead update is that this practice suits the way the retail firms operate when forecasting their sales and making their plans, creating also a rich set on which the ML classification method can be effectively trained, without exaggerating in terms of computational cost.

Finally, although we chose to train the classification method so that the average accuracy of CHF is minimized across the entire hierarchy, this objective can be easily adjusted in order for CHF to provide more accurate forecasts for a specific level of interest, as demonstrated in Appendix C. This choice depends on the decision-makers and can vary based on their focus and objectives. However, we do believe that our choice to optimize forecasting accuracy across the entire hierarchy, weighting equally all hierarchical levels, is realistic when dealing with demand forecasting and supply chain management given that in such settings each level supports very different, yet equally important decisions. A similar weighting scheme was adopted in the latest M competition, M5 (Makridakis et al. 2020), whose objective was to produce the most accurate point forecasts for 42,840 time series that represent the hierarchical unit sales of ten Walmart stores.

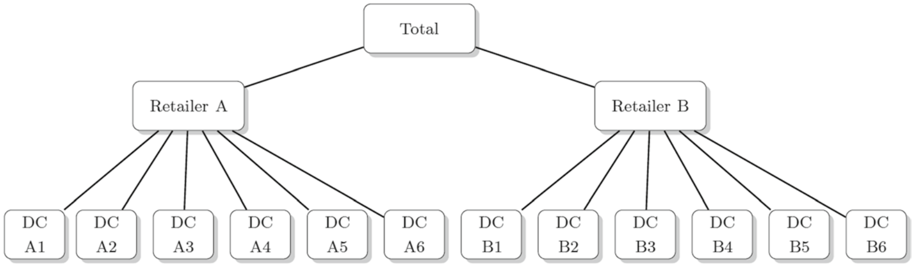

Fig. 3 Hierarchical structure of the time series included in the examined data set, representing the sales of 55 food products sold in Australia

Table 2 Number of time series per hierarchical level in the examined data set, representing the sales of 55 food products sold in Australia

Hierarchical level	Number of series
Level 0	1
Level 1	2
Level 2	12
Total	15

4 Data and experimental setup

4.1 Data

Although HF is relevant in many applications, such as energy (Spiliotis et al., 2020b) and tourism (Kourentzes and Athanasopoulos, 2019), it is most commonly found in the retail industry where SKU demand can be grouped based on location and product-related information. Therefore, the primary data set used in this study for empirically evaluating the accuracy of the proposed HF approach involves the sales and prices recorded for 55 hierarchies, corresponding to 55 fast-moving consumer goods of a food manufacturing company sold in various locations in Australia. Although the exact labels of the products are unknown to us, the products include breakfast cereals, long-life milk, and other breakfast products.

The hierarchical structure is the same for all 55 products of the data set and is depicted in Fig. 3. The number of series at each hierarchical level is provided in Table 2. As seen, each hierarchy consists of three levels, where the top level (level 0) represents total product sales, the middle level (level 1) the product sales recorded for 2 major retailer companies, and the bottom level (level 2) the way the product sales are disaggregated across 12 distribution centers (DC), 6 per retailer company, located in different states of Australia. Thus, each hierarchy includes 15 time series, each containing 120 weeks of observations, spanning from September 2016 to December 2018.

Figure 4 presents the hierarchical time series of an indicative product of the data set. Observe that sales may experience spikes at different levels of the hierarchy, i.e. different levels of the supply chain. These spikes correspond to promotional periods, and their frequency and size vary significantly for different products. Moreover, different nodes in the hierarchy may experience spikes of different extent. Finally, since each hierarchy corresponds to a different product, the series of the data set display different

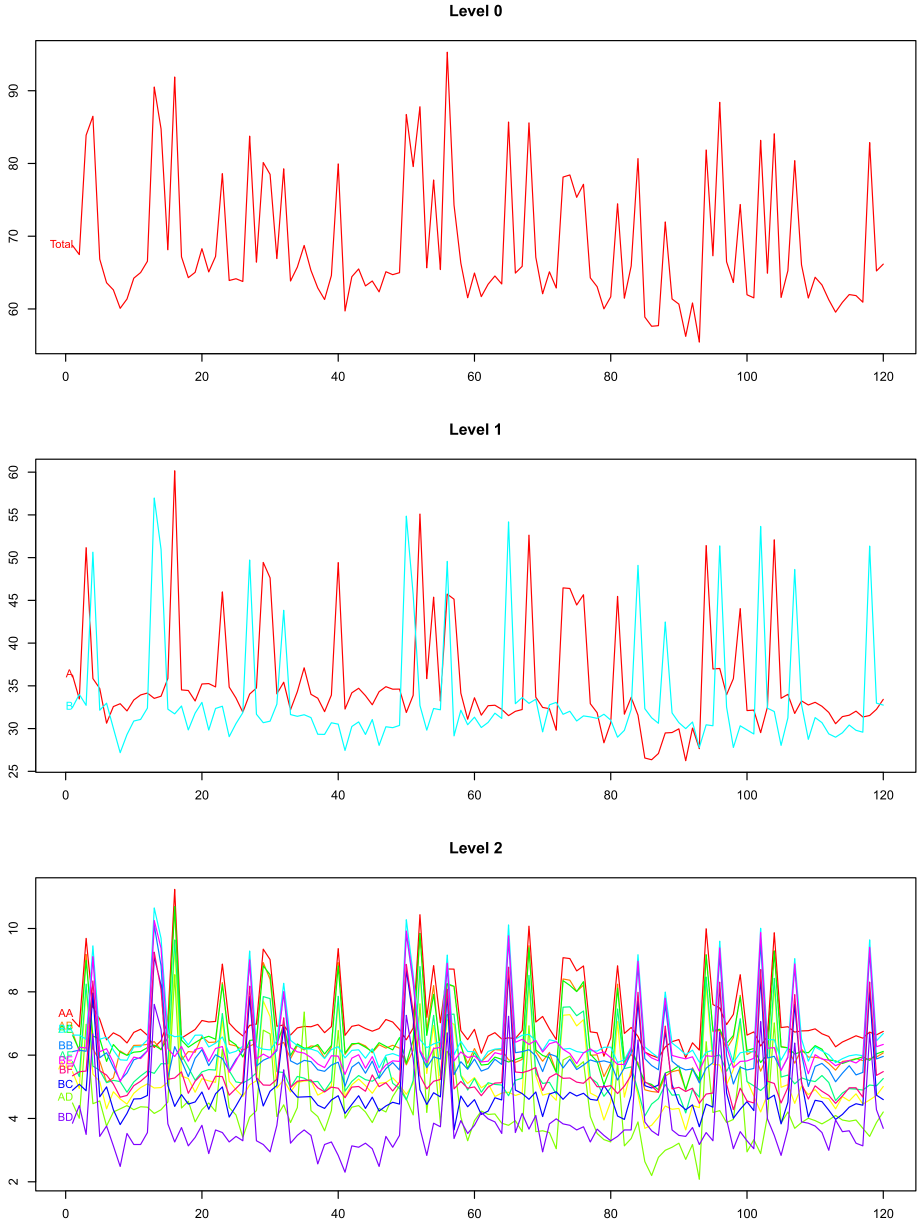

Fig. 4 Hierarchical sales of an indicative product of the examined data set. Level 0 represents total product sales, level 1 the product sales recorded for 2 major retailer companies, and level 2 the product sales across 12 distribution centers

strengths of trend and seasonality, volume, and entropy. For example, some series depict seasonality and small volume of sales, whereas some other series report large volume of sales and entropy, making them more volatile and difficult to forecast. As such, the data set represents a diverse set of demand patterns which are affected under the presence of promotions, i.e. price changes, among others.

In order to provide more empirical evidence regarding the effectiveness of the proposed approach over both standard and state-of-the-art HF methods on a diverse range of data, we have considered two more data sets in addition to the *Sales* data presented above. These data sets have different properties, thus giving a more diverse set of data for our empirical study. The first additional data set, to be called the *Tourism* data set, provides information on the Australian domestic tourism demand, measured as the number of overnight trips Australians spend away from home and disaggregated by state and region. The data set consists of 85 series, namely the total demand (level 0; 1 series), the demand per state (level 1; 8 series), and the demand per region (level 2; 76 series). The series are quarterly and span from q1-1998 to q4-2017 (80 periods). The second data set, to be called the *Prison* data set, provides information on the size of the prison population in Australia, disaggregated by state and gender. The data set consists of 25 series, namely the total prison population (level 0; 1 series), the prison population per state (level 1; 8 series), and the prison population of males and females per state (level 2; 16 series). The series are quarterly and span from q1-2005 to q4-2016 (48 periods). Each data set demonstrates a different structure, with its series being also characterized by different seasonal and trend patterns, randomness, lengths, and autocorrelation. Moreover, for each case, different forecasting horizons and optimization criteria have been used for employing CHF. The results of these experiments are summarized in Appendices 4.1 and 4.2.

4.2 Experimental setup

Considering that the examined data set involves products with sales which are highly affected by promotions, we produce base forecasts for all the series of the 55 hierarchies using a regression model with ARMA errors (Reg-ARMA), where product prices are used as a regressor variable. We choose Reg-ARMA for two reasons: First, it is a powerful method that can embody explanatory variables into the model and has been successfully implemented in various forecasting tasks (Abolghasemi et al., 2019; Abolghasemi et al., 2020). Second, it is dynamic in nature as its ARIMA component can account for unexplained variations, making it a desirable choice for forecasting sales time series that are impacted by promotions. Reg-ARMA can effectively capture sales both during promotional and non-promotional periods as price inherently carries the impact of promotions and, therefore, explains sufficiently the corresponding variations in sales. Reg-ARMA model is implemented using the *forecast* package for Hyndman et al. (2020a).

To evaluate the accuracy of the proposed HF approach both in terms of median and mean approximation (Kolassa, 2016), we consider two measures, namely the mean absolute scaled error (MASE) and the root mean sum of squared scaled error (RMSSE), respectively. The measures are calculated as follows:

$$\text{MASE} = \frac{n-1}{h}\frac{\sum_{t=n+1}^{n+h}|y_t - f_t|}{\sum_{t=2}^{n}|y_t - y_{t-1}|},$$

$$\text{RMSSE} = \sqrt{\frac{n-1}{h}\frac{\sum_{t=n+1}^{n+h}(y_t - f_t)^2}{\sum_{t=2}^{n}(y_t - y_{t-1})^2}},$$

where y_t and f_t are the observation and the forecast for period t, n is the sample size (observations used for training the forecasting model), and h is the forecasting horizon. Smaller

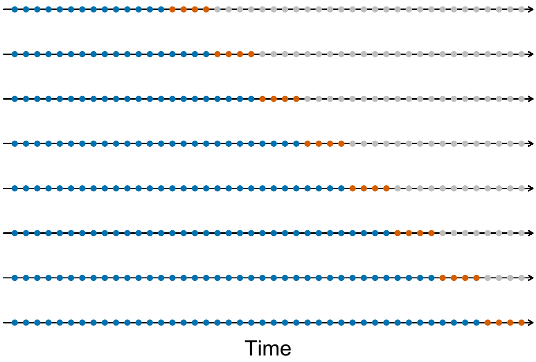

Fig. 5 Time series rolling origin approach for training, evaluating, and testing forecasting models. The blue circles represent the training set and the orange circles indicate the test set on each round

MASE and RMSSE values suggest more accurate forecasts. Note that both measures are scale-independent, thus allowing us to average the results across series.

Once the base forecasts are produced, we use the BU, TD, and COM methods to reconcile them across the three levels of the hierarchy. These baseline methods are used for benchmarking the proposed HF approach as they have been successfully applied in numerous applications and are considered standards in the area of HF (Abolghasemi et al., 2019; Hyndman et al., 2011; Hyndman et al., 2016). We also use the CHF method to select which of the three benchmarks is more suitable for forecasting each hierarchy.

In order to fit and evaluate the CHF method, we split the original data set into a training and test set. Specifically, we used the first 26 weeks of data to initially train the Reg-ARMA model and the following 58 periods to produce 4-step-ahead base forecasts on a rolling origin basis (Tashman, 2000). We considered 26 weeks of observations as the initial set to provide enough observations for training the forecasting models and then generated 4-step-ahead forecasts since this horizon (one month) is often enough in practice for operational planning on a weekly basis. Moreover, this creates a sufficient number of observations for training the classification model. We state that the initial number of observations and the number of forecast steps can be chosen differently, but one should generally consider a high enough number of observations for training both the baseline forecasting models and the classifiers.

Figure 5 depicts the rolling origin approach that we have used for training and testing the forecasting models. The blue circles represent our training set, while the orange circles indicate the test set on each round. Suppose a particular product with all 15 time series. At each iteration, and for each series, we produce 4-step-ahead forecasts and then roll the forecasting origin by four periods, i.e. we add four observations to the training data set and proceed by forecasting the following four periods. Each time that the forecast origin was updated, the set used for fitting the Reg-ARMA model was accordingly extended so that the base forecasts produced were appropriately adjusted. Moreover, on each step, the BU, TD, and COM methods were applied to reconcile the base forecasts and identify the most accurate alternative according to MASE. We consider the first 26 weeks as the initial training data set and repeat this process until the end of evaluation set, i.e. period 84, for each set of hierarchical time series. In this respect, a total of 14 accuracy measurements (average accuracy of 4-step-ahead forecasts over 58 weeks) × 55 hierarchies × 15 series = $11,550$ evaluations were recorded. We then summarized the results across the hierarchy and constructed a data set of 14 evaluations × 55 hierarchies = 770 observations that was used for training the XGB classification method. The remaining 36 weeks of data were used as a test set to evaluate the actual accuracy of the proposed approach, again on a rolling origin fashion. Thus, our evaluation is based on a total of 9 accuracy measurements (average accuracy of 4-step-ahead forecasts over 36 weeks) $\times$ 55 hierarchies $\times$ 15 series $= 7,425$ observations. *Note that XGB is retrained each time that we move across the window, with the values of the time series features being accordingly updated.*

5 Empirical results and discussion

Table 3 displays the forecasting accuracy (MASE and RMSSE) of the three HF methods considered in this study as benchmarks as well as classes for training the CHF algorithm in the retail data set presented in Sect. 4.1. The accuracy is reported both per hierarchical level and on average (arithmetic mean of the three levels), while CHF is implemented using the XGB classifier, as described in Sect. 4.2. Note that, as explained in Sect. 3, the average accuracy of the three hierarchical levels, as measured by MASE, is used for determining the training labels of the classifiers.

The results indicate that, on average, CHF is the best HF method according to both accuracy measures used. Specifically, CHF provides about 5%, 9%, and 2% more accurate forecasts than BU, TD, and COM, respectively, indicating that the XGB method has effectively managed to classify the hierarchies based on the features that their series display.

The improvements are similar if not better for the middle and bottom levels of the hierarchy. However, CHF fails to outperform TD and COM for the top hierarchical level, being about 6% and 2% less accurate, despite being still 8% better than the BU method. This finding confirms our initial claim that, depending on the hierarchical level of interest, different HF methods may be more suitable. In this study, we focused on the average accuracy of the hierarchical levels and optimized CHF with such an objective. However, as described in Sect. 3, different objectives could be considered in order to explicitly optimize the top, middle or bottom level of the hierarchy. In Appendix C we have examined such objectives and evaluated the respective performance of the CHF method. For instance, the results of Table 5 suggest that when XGB is optimized in terms of MASE and with respect to the top level, the forecast error of CHF is reduced from 0.466 (XGB-Avg) to 0.449 (XGB-L0), i.e., by 4% compared to the optimization criteria currently used.

Table 3 Forecasting performance of HF methods in terms of MASE and RMSSE over the test set

HF Method	Level 0	Level 1	Level 2	Average
MASE				
BU	0.503	0.854	0.856	0.738
TD	**0.435**	0.987	0.907	0.776
COM	0.455	0.844	0.844	0.714
CHF	0.466	**0.820**	**0.828**	**0.705**
RMSSE				
BU	0.583	1.049	1.044	0.892
TD	**0.507**	1.180	1.100	0.929
COM	0.531	1.039	1.035	0.868
CHF	0.534	**1.014**	**1.006**	**0.851**

Best solutions in each category are indicated in bold

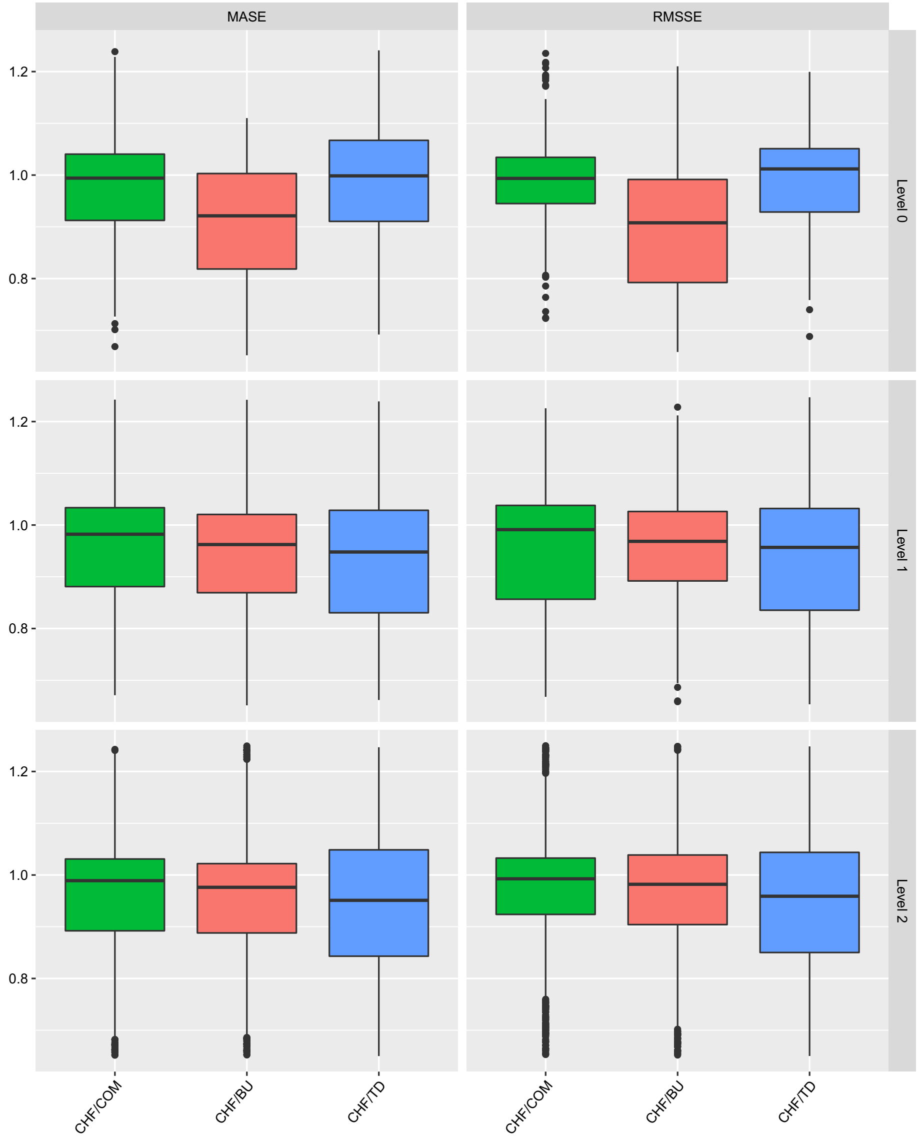

Fig. 6 The ratio of the forecasting accuracy of the CHF approach over the baseline HF methods across all 55 hierarchies included in the test set. The results are reported for each accuracy measure and hierarchical levels separately

In order to better evaluate the performance of the proposed approach, we proceed by investigating the distribution of the error ratios reported between CHF and the three benchmark methods examined in our study across all the 55 hierarchies of the data set. The results, presented per hierarchical level and accuracy measure, are visualized in Fig. 6 where box plots are used to display the minimum, 1st quantile, median, 3rd quantile, and maximum values of the ratios, as well as any possible outliers. Values lower than unity indicate that CHF provides more accurate forecasts and vice versa. As

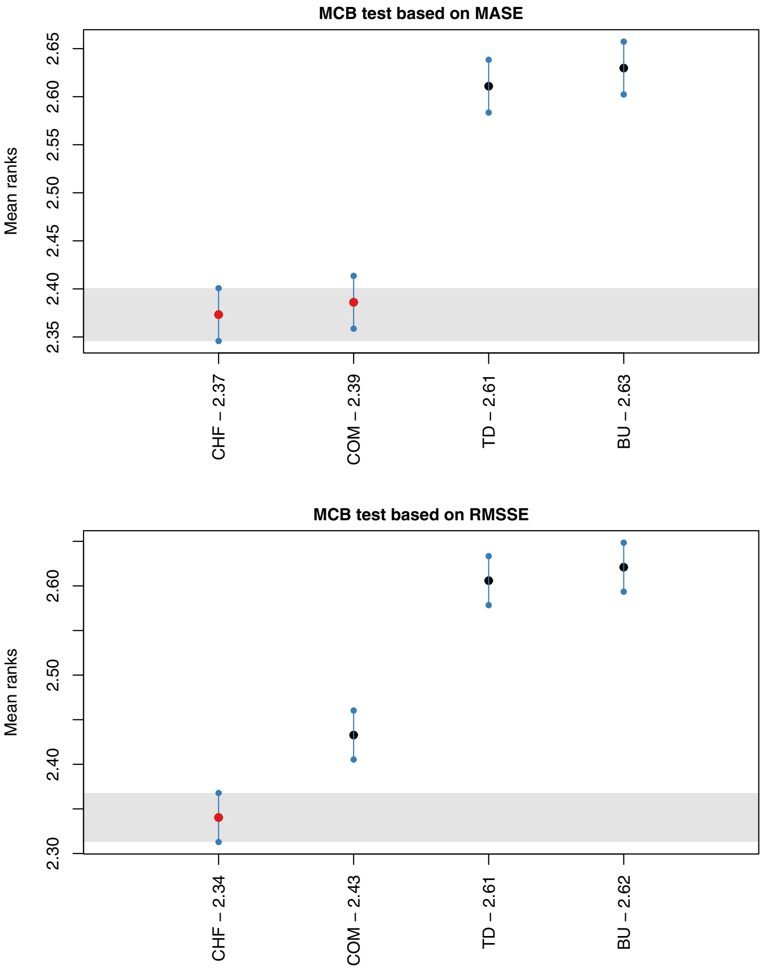

Fig. 7 MCB test conducted on the HF methods examined in this study using the test set of the sales time series data. MASE and RMSSE are used for computing the ranks and a 95% confidence level is considered. The ranks are computed considering all the series of the 55 hierarchies

observed, in most of the cases, CHF outperforms the rest of the HF methods, having a median ratio value lower than unity. The only exception is when forecasting level 0 and using RMSSE for measuring forecasting accuracy, where the TD method tends to provide superior forecasts. Thus, we conclude that CHF does not only provide the most accurate forecasts on average across all the 55 hierarchies, but also the most accurate forecasts across the individual ones.

To validate this finding, we also examine the significance of the differences reported between the various HF methods using the multiple comparisons with the best (MCB) test, as proposed by (Koning et al., 2005). According to MCB, the methods are first ranked based on the accuracy they display for each series of the hierarchy and then their average ranks are compared considering a critical difference, $r_{\alpha,K,N}$, as follows:

Table 4 Performance of the XGB classification method

ML classifier	Class	Precision	Recall	F_1
XGB (MASE)	BU	0.31	0.44	0.36
	TD	0.42	0.23	0.30
	COM	0.42	0.48	0.45
XGB (RMSSE)	BU	0.29	0.31	0.30
	TD	0.41	0.34	0.38
	COM	0.39	0.44	0.41

$$r_{\alpha,K,N} = q_a \times \sqrt{\frac{K * (K+1)}{12N}}, \tag{2}$$

where N is the number of the time series, K is the number of the examined HF methods, and q_a is the quantile of the confidence interval. In our case, where α is set equal to 0.05 (95% confidence), q_a takes a value of 3.219. Accordingly, K is equal to 4 (TD, BU, COM, CHF) and N is equal to 7,425.

The results of the MCB test are presented in Fig. 7. If the intervals of two methods do not overlap, this indicates a statistically different performance. Thus, methods that do not overlap with the gray zone of Fig. 7 are significantly worse than the best, and vice versa. As seen, our results indicate that CHF provides significantly better forecasts than the rest of the HF methods, both in terms of MASE and RMSSE. Moreover, we find that CHF is followed by COM and then by TD and BU. Interestingly, the performance of TD is not significantly different than the one of BU according to RMSSE or MASE. In this regard, we conclude that CHF performs better than the state-of-the-art HF methods found in the literature, being also significantly more accurate than the standards used for HF, such as BU and TD. As such, CHF can be effectively used for selecting the most appropriate HF method from a set of alternatives and improving the overall forecasting accuracy of various hierarchies.

We also investigate the performance of the XGB classifier in terms of precision, recall, and F_1 score. The precision metric measures the number of correct predictions in the total number of predictions made for each class. Recall (also known as sensitivity) informs us about the number of times that the classifier has successfully chosen the best HF method for each class. Finally, the F_1 score is the harmonic mean of precision and recall, computed by $F_1 = 2\frac{p*r}{p+r}$, and is used to combine the information provided by the other two metrics (Sokolova and Lapalme, 2009).

Before presenting the results, we note that the TD, BU, and COM methods have been identified as the most accurate HF method in 1860, 2745, and 2820 cases, respectively. This indicates that the data set used for training the classification method was equally populated, displaying a uniform probability distribution. Having a balanced training sample is important for our experiment since it facilitates the training of the ML algorithm (enough observations from each class are available and biases can be effectively mitigated) and provides more opportunities for accuracy improvements (if no HF method is dominant, selecting between different HF methods becomes promising). The opposite is expected to be true for highly imbalanced data sets: Given a dominant HF method, even if the classifier is effectively trained, little room for improvements in terms of forecasting accuracy will be available.

The performance of the XGB classification method is presented in Table 4. According to the precision metric, XGB manages to select the COM methods more accurately

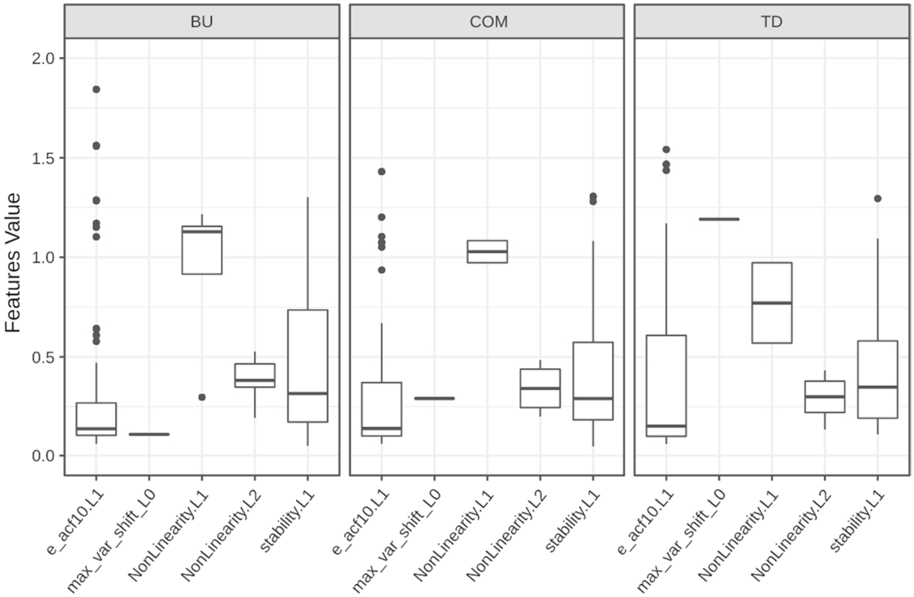

Fig. 8 The five features most frequently used by the XGB classification method. The features are displayed for each HF method, i.e. when selected by the classifier for performing the reconciliation, separately

but finds it difficult to make appropriate selections when the BU or TD method is preferable according to the MASE metric. This indicates that, although XGB identifies the conditions under which the COM method is preferable, the opposite is not true. We provide the details of confusion matrices in Appendix D.

As a final step in our analysis, we investigate the significance of the time series features used by the XGB classifier, i.e. the number of times that each feature was considered by the method for making a prediction. We compute 32 features for each time series and select the top 5 of them when reconciling with different HF methods. Figure 8 presents the features that are most frequently selected by the classifier. As seen, the *non-linearity* at levels 1 and 2, *stability* at level 1, *e-acf10* at level 1, and *max-var-shift* at level 0 are the most frequently used features and, therefore, the most critical variables for selecting a HF method. The distribution of these features vary for different selected HF methods. Non-linearity at levels 1 and 2 is among of the strongest features in our data set and stability at level 1 also plays an important role. We believe this is because promotions are frequently affecting the sales of the products strongly, thus changing the volatility of their demand both over promotional and non-promotional periods (Abolghasemi et al. 2020). Maximum variance shift at level 0 is another feature that is frequently selected. This may attribute to the sudden changes and spikes caused to sales data set by promotions. TD is selected for higher values of this metric followed by COM and BU. Finally, *e-acf10*, which contains the sum of squared values of the first 10 autocorrelation coefficient of the error terms of series at level 1, is also among the top selected features. One possible explanation is because sales are being affected by promotions and therefore depict high levels of variations during promotion periods. Therefore, even after fitting a forecasting model, there will still be some correlation in the remainder term.

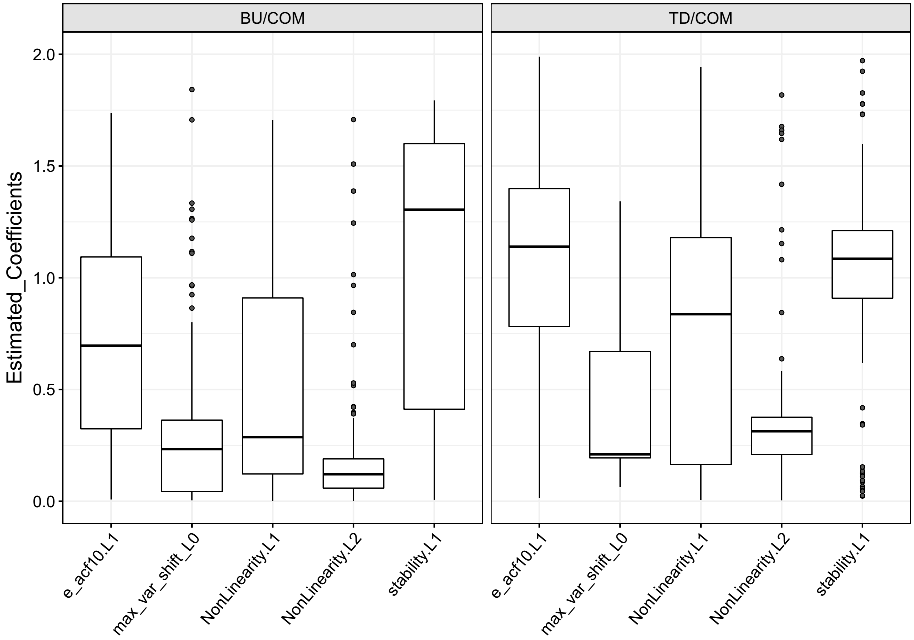

Fig. 9 The coefficients of the top 5 features estimated by the LR classification method

In order to shed more light on the importance of the time series characteristics in selecting a HF method, we investigated the performance of a LR classifier, which is a statistical and more interpretable classification model, and focused on the distributions of the estimated coefficients of the multinomial logistics. Figure 9 shows the boxplots of the estimated coefficients for the top 5 features of the series, as determined earlier by XGB. We consider the COM method as the reference group in the LR model and show two groups of boxplots including BU/COM and TD/COM coefficient ratios. Since the parameter estimates are relative to the reference group, the estimated values on the y-axis show how much the log-odds for the corresponding method is expected to change for a unit change in the time series characteristics if all the other characteristics in the model are held constant. For example, the odds of selecting BU over COM increase by less than 0.24% when maximum variance shift at level 0 (*max-var-shift-L0*) increases by one unit. Interestingly, when the stability at level 1 (*stability.L1*) increases by one unit, the odds of selecting BU over COM increase by more than 1%, on average. The results are different for selecting TD over COM. As we can see, the odds of selecting TD to COM increase the most when the sum of the first 10 auto-correlations at level 1 (*e-acf-10.L1*) increases by one unit.

We also trained a decision tree model with data for all series and hierarchies to visualise the selection process implemented by a single decision tree. We included the decision tree in Appendix D.

Note that when conducting this experiment we considered another training set-up for the classifiers where, apart from time series features, we also provided information about the correlation of the series, both across levels and within each level separately, similar to (Nenova and May, 2016). The results were similar to those reported in Table 3, and therefore we decided to exclude those features from our models for reasons of simplicity. In another experiment, we implemented the model of (Nenova and May, 2016) on our data set where we used series correlation instead of time series characteristics as input predictors. We report the details and results in Appendix E.

Another extension of the proposed approach that could be also examined in a future study is that it focuses on selecting the most appropriate hierarchical forecasting method per hierarchy. However, numerous empirical studies have shown that combining forecasts from multiple forecasting methods can improve forecasting accuracy (Makridakis et al., 2020; Lemke and Gabrys, 2010; Atiya, 2020). Thus, replacing classifiers with other methods that would combine various HF methods using appropriate weights becomes a promising alternative to CHF. Simple, equal-weighted combinations of standard HF methods have already been proven useful under some settings (Mircetic et al., 2021; Abouarghoub et al., 2018), while feature-based forecast model averaging has demonstrated its potential to generate robust and accurate forecasts (Montero-Manso et al., 2020).

6 Conclusion

This paper introduced conditional hierarchical forecasting, a dynamic approach for effectively selecting the most accurate method for reconciling incoherent hierarchical forecasts. Inspired by the work done in the area of forecasting model selection and the advances reported in the field of machine learning, the proposed approach computes various features for the time series of the examined hierarchy and relates their values to the forecasting accuracy achieved by different hierarchical forecasting methods, such as bottom-up, top-down, and combination methods, using an appropriate classification method. Based on the lessons learned, and depending on the characteristics of time series in the hierarchy, the most suitable hierarchical forecasting method can be chosen and used to enhance overall forecasting performance.

We exploited various time series features at different levels of the hierarchy that represent their behavior, and trained an extreme gradient boosting classification model to choose the most appropriate type of hierarchical forecasting method for a hierarchical time series with the selected features. The accuracy of the proposed approach was evaluated using a large data set coming from the retail industry and compared to that of three popular hierarchical forecasting methods. Our results indicate that conditional hierarchical forecasting can produce significantly more accurate forecasts than the benchmarks considered, especially at lower hierarchical levels. Thus, we suggest that, when dealing with hierarchical forecasting applications, selection should be expanded from forecasting model to reconciliation methods as well. We further validated our approach by experimenting on two additional and diverse datasets including tourism and prison datasets. These datasets have different properties and confirm our findings that the best reconciliation method can be selected as per time series characteristics and structure of the hierarchy. These factors may also impact the performance of a classifier.

Undoubtedly, our study displays some limitations that are worth investigating in future endeavors. The forecasting performance of the conditional hierarchical forecasting algorithm depends on the classification performance of the models used for its implementation and the data used for their training. For instance, if the training data set available is highly imbalanced, i.e. one hierarchical forecasting method is dominant over others, this can diminish the performance of the classification models. The class-imbalance in the training set can be more severe if we consider a larger number of hierarchical forecasting methods, thus making the multi-class classification task more challenging. Developing an algorithm that can deal with these issues within the proposed framework could help improve further the overall performance of the proposed method. Another avenue for future research that seems a natural extension to our study is to use features of the hierarchy, e.g. correlations between series, number of levels, and number of series, as alternative inputs for selecting the most appropriate reconciliation method.

Appendix A: CHF Algorithm

The CHF algorithm, presented in detail in Sect. 3, is summarised below.

Algorithm Conditional Hierarchical Forecasting

```
1:  Off-line phase
2:  for $t = p$ to $r$ with a step of $h$ do
3:      Create a train and test set by splitting the available in-sample data of size $r$. The train set includes the first $p$ observations and the test set includes the following $h$ observations, $p+1:p+h$, equal in number with the forecasting horizon considered, $h$.
4:      Fit a forecasting method of choice to the train set and produce $h$-step-ahead forecasts.
5:      Reconcile the base forecasts produced in step 4 with different HF methods of choice.
6:      Calculate the accuracy of the $h$-step-ahead forecasts produced in step 5 by the baseline HF methods considered using a measure of choice. Accuracy can be computed across all hierarchical levels or a particular one, depending on the objective of the forecasting task.
7:      Compute a variety of time series features ($z$ in number) for the $m$ series of the hierarchy.
8:  Create a train set for a ML classification method of choice. The train set of the classification method includes the average values of the time series features considered in step 7, computed across all the series of each hierarchical level separately. Thus, a total of $k \times z$ independent numerical variables are provided as input to the classifier. The target variable is the most accurate HF method, as determined in step 6, and is provided to the classifier as a categorical variable.
9:  Train the ML classification method using the train set developed in step 8.
10: On-line phase
11: for $t = r+1$ to $n$ with a step of $h$ do
12:     Compute the time series features considered in step 7 for all the $m$ series of the hierarchy up to observation $t$. Then, compute the average value of these features per hierarchical level, as done in step 8.
13:     Use the classification method trained in step 9 and the input data constructed in step 12 to predict the class of the examined hierarchy, i.e., the most accurate HF method from the ones considered in step 5.
14:     Produce base forecasts for the following $h$ periods, $t+1:t+h$, and reconcile these forecasts using the HF method predicted in step 13.
```

Appendix B: Time series characteristics used in this study

In this appendix, we provide a brief explanation of the time series features considered in the present study for developing the classification models used within the CHF approach. Our list is inspired by recent studies that have successfully used time series features to develop meta-learning forecasting algorithms (e.g. for model selection and combination, Montero-Manso et al., 2020), being tailored however for the particular requirements of the conditional hierarchical forecasting task.

1. *Entropy*: Measures the "forecastability" of a time series. Lower values of entropy suggest higher signal to noise ratios that make a series easier to forecast (Garland et al., 2014; Goerg , 2013; Liu et al., 2016). Entropy is calculated as shown in Eq. (3),

$$\text{entropy} = -\int_{-\pi}^{\pi} \hat{f}(\lambda)\log(\hat{f}(\lambda))d\lambda, \tag{3}$$

   where $\hat{f}(\lambda)$ is the spectral density of the data, describing the strength of a time series as a function of frequency λ.
2. *Lumpiness*: Measures the variance of the variances of non-overlapping windows in a series.
3. *Stability*: Measures the variance of the mean of non-overlapping windows in a series.
4. *Hurst*: Measures the long-term memory of a time series. Hurst is equal to 0.5 plus the maximum likelihood estimate of the fractional differencing order d, where d is the degree of first differencing after fitting an autoregressive fractionally integrated moving average model to the time series.
5. *Seasonal-period*: Indicates the number of seasonal periods of a series. If a series is not seasonal, the metric takes a value of 1.
6. *Easonal-strength*: Time series depict seasonality when they exhibit a pattern that is repetitive and it is caused by seasonal factors. As such, time series with a fixed seasonality will display significant autocorrelation at fixed seasonal lags. Suppose that S_t, T_t, and E_t represent the seasonality, trend, and error components of a time series Y so that $Y_t = S_t + T_t + e_t$. Based on this assumption, we can detrend $X_t = Y_t - T_t$ and deseasonlize a time series $Z_t = Y_t - S_t$. Similarly, we can subtract the trend and seasonality from the series to compute the remainder (error term) of the underlying decomposition approach, $e_t = Y_t - T_t - S_t$. The strength of seasonality can then be computed as in Eq. (4).

$$\text{Seasonality strength} = 1 - \frac{\text{Var}(e_t)}{\text{Var}(X_t)} \tag{4}$$

7. *Trend*: Trend indicates a long-term change in the mean of a time series. We measure the strength of the trend using Eq. (5).

$$\text{Trend strength} = 1 - \frac{\text{Var}(e_t)}{\text{Var}(Z_t)} \tag{5}$$

8. *Curvature*: Measures the curvature of a time series. The measure is calculated based on the coefficients of an orthogonal quadratic regression.
9. *e-acf1*: Measures the first autocorrelation coefficient after calculating the autocorrelation of the remainder of series, e_t.

10. *e-acf10*: It is the sum of the squares of the first ten autocorrelation coefficients after calculating the autocorrelation of the remainder of series, e_t.
11. *x-acf1*: Measures the first order autocorrelation of the series.
12. *x-acf10*: It is the sum of the squares of the first ten autocorrelation coefficients of the series.
13. *diff1-acf1*: Measures the first order autocorrelation of the differenced series.
14. *diff1-acf10*: It is the sum of the squares of the first ten autocorrelation coefficients of the differenced series.
15. *diff2-acf1*: Measures the first order autocorrelation of the twice differenced series.
16. *diff2-acf10*: It is the sum of the squares of the first ten autocorrelation coefficients of the twice differenced series.
17. *seas-acf1*: Measures the autocorrelation of the seasonality component of the series.
18. *x-pacf5*: It is the sum of the squares of the first five partial autocorrelation coefficients of the series.
19. *diff1x-pacf5*: It is the sum of the squares of the first five partial autocorrelation coefficients of the differenced series.
20. *diff2x-pacf5*: It is the sum of the squares of the first five partial autocorrelation coefficients of the twice differenced series.
21. *seas-pacf*: Measures the partial autocorrelation of the seasonality component of the series.
22. *Linearity*: Measures the linearity of a time series. It is calculated based on the coefficients of an orthogonal quadratic regression.
23. *Non-linearity*: Measures the non-linearity of a time series. It is calculated using a modified version of the Teräsvirta's non-linearity test as described in (Hyndman et al. , 2019).
24. *max-var-shift*: Measures the largest variance shift between to consecutive windows in a time series.
25. *max-kl-shift*: Measures the largest Kulback-Leibler divergence between to consecutive windows in a time series.
26. *fluctanal-prop-r1*: Measures the fluctuation of a series. It fits a polynomial of order 1 and then returns the range.
27. *unitroot-kpss*: A time series is stationary if its mean, variance, and autocovariance do not depend on time. We use Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests to check whether the time series is stationary or not. KPSS test uses a null hypothesis that an observable time series is stationary against the alternative of a unit root (Kwiatkowski et al., 1992).
28. *arch-acf*: It is the sum of squares of the first 12 autocorrelation values of a pre-whitened time series.
29. *garch-acf*: It is the sum of squares of the first 12 autocorrelation values of the residuals after fitting a GARCH(1,1) model to a pre-whitened time series.
30. *arch-r2*: It is the R^2 value of an AR model applied to a pre-whitened time series.
31. *garch-r2*: It is the R^2 value of an AR model applied to the residuals after fitting a GARCH(1,1) model to a pre-whitened time series.
32. *arch-test*: Measures autoregressive conditional heteroscedasticity (ARCH). This value is the R^2 value of an autoregressive model of order specified as lags applied to x^2.

Appendix C: Forecasting performance of CHF when additional data sets, optimization criteria, and classification methods are considered

In this appendix we summarize the forecasting performance of the three baseline HF methods (BU, TD, and COM) considered in the present study in terms of MASE and RMSSE, as well as the proposed CHF one when LR, LDA, DT, RF, SVM, XGB, or NN classification models are used for its implementation. We do so in order to provide more evidence regarding the impact of the classification method used within the proposed approach. Moreover, the results are presented for four different optimization criteria depending on the particular objective of the classification task, i.e., which forecasts should be considered as "optimal". Specifically, we train the classifiers so that the reconciled forecasts produced by the CHF method are optimal in terms of the accuracy measured at (i) level 0, (ii) level 1, (iii) level 2, or (iv) all levels (average performance across levels 0, 1, and 2). We perform such an analysis since, depending on the decisions the forecasts opt to support, different cross-sectional levels may be more relevant, meaning that CHF should be flexible enough to be adapted accordingly.

In addition to the *Sales* data set presented in the main part of the paper, we consider two more data sets, namely the *Prison* and the *Tourism* ones Hyndman and Athanasopoulos (2021). By doing so we provide more empirical evidence regarding the effectiveness of the proposed method when hierarchies of different structures, series of different frequencies, lengths, and characteristics, or forecasting horizons are considered. In both cases, the base forecasts were produced using ExponenTial Smoothing (ETS, Hyndman, 2002), as implemented in the *forecast* package for Hyndman 2020a.

LR and LDA were implemented using the *nnet* and *MASS* packages for R, respectively [nnetR,MASSr]. DT was implemented using *rpart* for R [rpartR], using a complexity parameter of 0.01. RF was implemented using the *randomForest* package for R Liaw and Wiener (2002). We set the number of trees equal to 150 and optimized the rest of its hyper-parameters using a grid search in a 5-fold cross-validation fashion. The minimization of the error rate was used as a loss function. The optimal number of nodes was selected between 2 and 10 with an interval of 1, while the minimum size of the terminal nodes was selected between 1 and 5 with an interval of 1 Probst et al. (2019). SVM was implemented using the *kernlab* package for R (Karatzoglou et al., 2004). We chose the Radial Basis kernel and optimized the cost of the constraint violation, *C*, using a grid search between 0 and 300 with a step of 10 and a 5-fold cross-validation. The minimization of the error rate was used as a loss function. NN was implemented using the *nnet* package for R [nnetR]. We considered 3 fully-connected hidden layers and a logistic activation function, leaving the rest of the hyper-parameters to the default values of the package as the size of the data sets available for training do not allow for extensive cross-validation.

Sales data set

This part of the appendix presents the results of the additional experiments conducted for the case of the 55 data sets of hierarchical sales series. Six different classification models are used in addition to XGB for implementing the CHF approach, while various levels of the hierarchy are used as targets for optimizing the results. Tables 5 and 6 summarize our findings in terms of MASE and RMSSE, respectively. The significance of the results is assessed using MCB tests, as shown in Figs. 10 and 11. Our results are in line with those of

Table 5 Forecasting performance of the HF methods examined in this study in terms of MASE for the *Sales* data set

HF Method	Level 0	Level 1	Level 2	Average
BU	0.503	0.854	0.856	0.738
TD	**0.435**	0.987	0.907	0.776
COM	0.455	0.844	0.844	0.714
LR-Avg	0.508	0.858	0.869	0.844
LDA-Avg	0.509	0.855	0.867	0.841
DT-Avg	0.495	0.855	0.866	0.840
RF-Avg	0.470	0.846	0.848	0.721
XGB-Avg	0.466	0.820	0.828	0.705
SVM-Avg	0.457	0.850	0.850	0.719
NN-Avg	0.495	0.855	0.866	0.840
LR-L0	0.502	0.912	0.902	0.877
LDA-L0	0.515	0.900	0.894	0.869
DT-L0	0.495	0.855	0.866	0.840
RF-L0	0.466	0.897	0.884	0.749
XGB-L0	0.449	0.889	0.857	0.732
SVM-L0	0.477	0.879	0.867	0.741
NN-L0	0.495	0.855	0.866	0.840
LR-L1	0.495	0.859	0.870	0.843
LDA-L1	0.494	0.857	0.869	0.842
DT-L1	0.495	0.855	0.866	0.840
RF-L1	0.465	0.852	0.851	0.723
XGB-L1	0.466	**0.815**	**0.828**	**0.703**
SVM-L1	0.464	0.861	0.858	0.728
NN-L1	0.495	0.855	0.866	0.840
LR-L2	0.502	0.885	0.883	0.857
LDA-L2	0.504	0.881	0.881	0.855
RF-L2	0.464	0.854	0.853	0.724
DT-L2	0.495	0.855	0.866	0.840
XGB-L2	0.468	0.840	0.840	0.716
SVM-L2	0.465	0.852	0.849	0.722
NN-L2	0.495	0.855	0.866	0.840

Best solutions in each category are indicated in bold

L0, L1, L2, and Avg indicate that the classification model used for implementing CHF is optimized with the objective of minimizing forecast error at level 0, level 1, level 2, and the average of all levels, respectively

Table 3, suggesting that CHF can result in superior forecasts compared to well-established HF methods when XGB is used for performing the classification. Moreover, we find that, more often than not, CHF can effectively adapt to the objective of the optimization process, thus being tailored to accurately forecast series at different hierarchical levels of interest. In addition, our results indicate that selecting an appropriate classification method that has the capacity to learn how to optimally link numerous time series features with forecasting performance, is critical for effectively implementing CHF. Interestingly, none of the

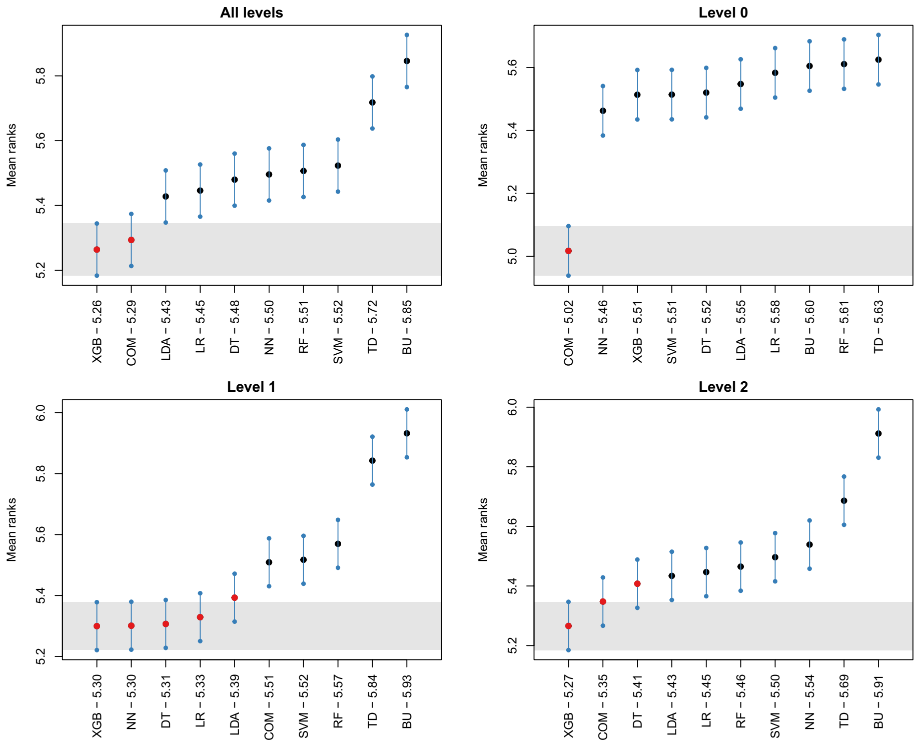

Fig. 10 MCB tests conducted on the HF methods examined in this study for the *Sales* data set. The results are presented for each optimization criterion separately, i.e., when the labels of the classification models are determined so that the forecasts produced are optimal in terms of level 0, level 1, level 2, or the average of all levels. In all cases, MASE is used for computing the ranks and a 95% confidence level is considered

additional classification methods considered performed equally well with XGB, being also outperformed by COM in most of the cases.

In order to better interpret our results and justify the superiority of the forecasts provided by XGB over the rest of the classification models utilized in this study for implementing CHF, we investigated the performance of the classifiers by analysing their confusion matrices. Table 7 compares the predictions made by the classification models with the true labels when the HF methods are determined so that the forecasts produced are optimal in terms of MASE for different hierarchical levels. We observe that XGB has managed to classify correctly more instances when the optimisation focused at level 1 and the average of all levels, while DT is the top-performing model when the optimisation is focused at level 0 and level 2. However, DT has misclassified more series to TD and BU when COM was the superior model, which inevitably contributed to its inferior forecast accuracy, on average. Similar conclusions can be drawn when RMSSE is used for measuring forecasting performance since, as shown in Table 8, XGB displays the highest classification accuracy in all cases, except at level 1. Again, DT has classified more series correctly at level 1 but the inferior performance in misclassifying COM has contributed to its lower accuracy, on average.

Table 6 Forecasting performance of the HF methods examined in this study in terms of RMSSE for the *Sales* data set

HF Method	Level 0	Level 1	Level 2	Average
BU	0.583	1.049	1.044	1.013
TD	**0.507**	1.180	1.100	1.071
COM	0.531	1.039	1.035	1.002
LR-Avg	0.547	1.064	1.046	1.015
LDA-Avg	0.549	1.059	1.043	1.012
DT-Avg	0.551	1.040	1.038	1.006
RF-Avg	0.592	1.072	1.072	1.016
XGB-Avg	0.534	**1.014**	**1.006**	**1.000**
SVM-Avg	0.594	1.059	1.065	1.008
NN-Avg	0.526	1.081	1.047	1.017
LR-L0	0.543	1.107	1.074	1.043
LDA-L0	0.546	1.090	1.067	1.035
DT-L0	0.561	1.049	1.044	1.026
RF-L0	0.544	1.105	1.076	1.049
XGB-L0	0.544	1.095	1.066	1.001
SVM-L0	0.531	1.098	1.067	1.049
NN-L0	0.541	1.087	1.067	1.014
LR-L1	0.544	1.049	1.041	1.009
LDA-L1	0.551	1.041	1.037	1.005
DT-L1	0.545	1.044	1.042	1.028
RF-L1	0.555	1.046	1.041	1.013
XGB-L1	0.537	1.016	1.009	1.001
SVM-L1	0.546	1.057	1.050	1.012
NN-L1	0.538	1.079	1.060	1.002
LR-L2	0.543	1.050	1.043	1.011
LDA-L2	0.544	1.044	1.041	1.008
DT-L2	0.544	1.043	1.043	1.010
RF-L2	0.538	1.041	1.039	1.010
XGB-L2	0.541	1.021	1.023	1.000
SVM-L2	0.545	1.056	1.044	1.009
NN-L2	0.543	1.085	1.064	1.003

Best solutions in each category are indicated in bold

L0, L1, L2, and Avg indicate that the classification model used for implementing CHF is optimized with the objective of minimizing forecast error at level 0, level 1, level 2, and the average of all levels, respectively

Appendix D: Variables importance

Finally, we trained a DT classifier on all 55 hierarchies of the data set to investigate the explicit conditions under which a particular HF method is more likely to be preferred over other alternatives. Figure 12 presents the structure of the DT and the rules defined by the model for performing the classification. While this model differs in terms of training from those originally considered in our study, in the sense that it is trained across all hierarchical series and not for each hierarchy separately, it sheds light on the selection process

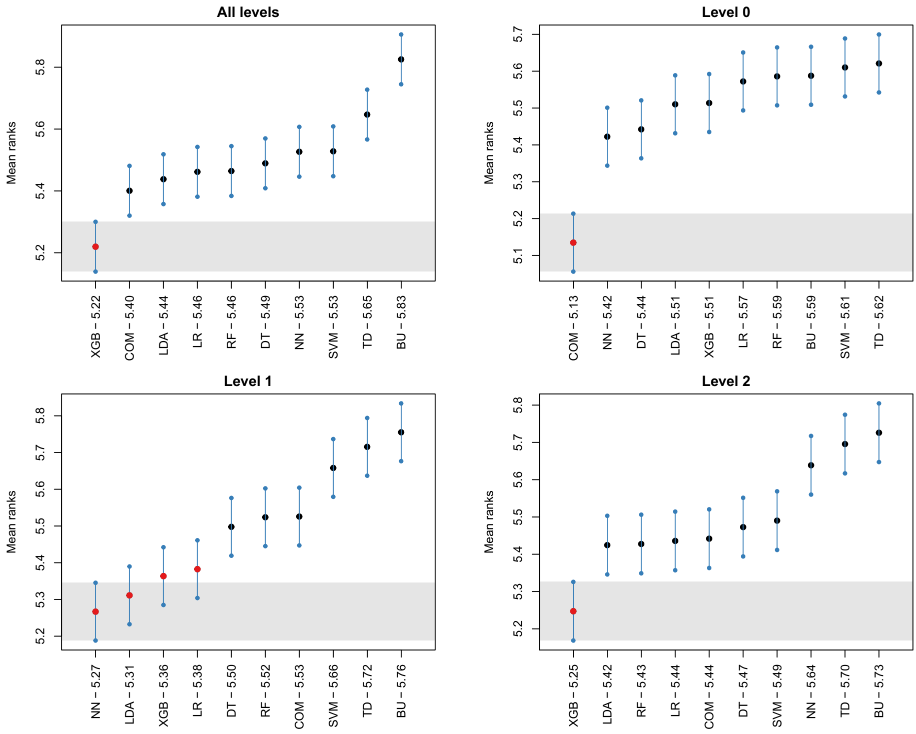

Fig. 11 MCB tests conducted on the HF methods examined in this study for the *Sales* data set. The results are presented for each optimization criterion separately, i.e., when the labels of the classification models are determined so that the forecasts produced are optimal in terms of level 0, level 1, level 2, or the average of all levels. In all cases, RMSSE is used for computing the ranks and a 95% confidence level is considered

performed by decision-tree-based classifiers. For example, we observe that when the series have higher values of correlation (*x-acf10.L2*≥ *0.75*), there is a richness of information at the bottom level and the bottom series are predictable, making BU is the most preferable HF method without necessarily using information from other levels. However, when the correlation at the bottom series is low but variance at the top level is high (*NonLinearity.L2<0.75 & max-var-shift-L0* ≥ *1.6*), then COM is the selected model. For smaller values of variance shift at the top level, TD, BU, and COM have been selected over 44%,19%, and 6% of the times, respectively. This indicates that when the variance shift at the top level is smaller and correlation at the bottom level is low, proportions will be a good representative to disaggregate them and the series at the top level is helpful for predicting those at the lower levels, enabling TD to perform better on average.

Prison data set

The *Prison* data set consists of 25 quarterly series, organized at three hierarchical levels (total, state, gender) and spanning over q1-2005 to q4-2016 (48 periods). Given that the series are relatively short but a reasonable amount of observations is required both for effectively training the baseline forecasting models and precisely computing the time series features, we start producing base forecasts after the 6th year and use the following

Table 7 Confusion matrices of the classification models considered for forecasting the series of the *Sales* data set when the HF methods (labels) are determined so that the forecasts produced are optimal in terms of MASE at level 0, level 1, level 2, or the average of all levels

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			Average		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
LR	BU	57	52	36	26	5	99	27	26	72	53	11	62
	TD	96	99	40	17	18	163	40	32	113	68	13	109
	COM	27	56	32	24	13	130	36	29	120	66	11	102
		Total correct: 188			Total correct: 174			Total correct: 179			Total correct: 168		
LDA	BU	58	54	33	24	6	100	30	25	70	53	13	60
	TD	98	105	32	14	19	165	35	31	119	69	13	108
	COM	27	58	30	20	11	136	38	27	120	68	8	103
		Total correct: 193			Total correct: 179			Total correct: 181			Total correct: 169		
DT	BU	68	58	19	40	53	37	37	56	32	68	19	39
	TD	102	114	19	39	100	59	30	106	49	73	48	69
	COM	43	58	14	53	69	45	51	88	46	82	38	59
		Total correct: **196**			Total correct: 185			Total correct: **189**			Total correct: 175		
RF	BU	58	47	40	30	30	70	38	30	57	39	13	74
	TD	98	74	63	22	59	117	38	41	106	45	36	109
	COM	37	41	37	36	40	91	49	43	93	49	27	103
		Total correct: 169			Total correct: 180			Total correct: 172			Total correct: 178		
XGB	BU	74	59	12	34	29	67	46	30	49	56	24	46
	TD	113	99	23	16	78	104	59	59	67	71	44	75
	COM	40	57	18	28	53	86	53	57	75	56	37	86
		Total correct: 191			Total correct: **198**			Total correct: 180			Total correct: **186**		

Table 7 (continued)

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			Average		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
SVM	BU	67	43	35	38	24	68	36	27	62	30	15	81
	TD	108	88	39	48	46	104	44	39	102	40	27	123
	COM	46	40	29	39	31	97	32	46	107	37	27	115
		Total correct: 184			Total correct: 181			Total correct: 182			Total correct: 172		
NN	BU	58	54	33	26	37	67	31	38	56	39	40	47
	TD	79	96	60	34	72	92	54	54	77	52	67	71
	COM	41	46	28	28	41	98	56	47	82	48	56	75
		Total correct: 182			Total correct: 196			Total correct: 167			Total correct: 181		

Best solutions in each category are indicated in bold

The rows correspond to the true optimal HF methods while the columns to the predictions made by the various classification models in the test set (rolling origin evaluation)

Table 8 Confusion matrices of the classification models considered for forecasting the series of the *Sales* data set when the HF methods (labels) are determined so that the forecasts produced are optimal in terms of RMSSE at level 0, level 1, level 2, or the average of all levels

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			All levels		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
LR	BU	44	54	45	25	33	64	25	21	70	26	36	57
	TD	61	119	65	15	68	119	30	50	114	29	67	102
	COM	25	50	32	38	49	84	28	46	111	34	54	90
		Total correct: 195			Total correct: 177			Total correct: 186			Total correct: 183		
LDA	BU	43	61	39	30	32	60	33	22	61	30	36	53
	TD	61	122	62	19	71	112	31	52	111	32	61	105
	COM	21	51	35	45	51	75	35	46	104	42	50	86
		Total correct: 200			Total correct: 176			Total correct: 189			Total correct: 177		
DT	BU	81	51	11	48	43	31	33	28	55	47	41	31
	TD	120	111	14	34	96	72	29	63	102	43	64	91
	COM	53	42	12	57	71	43	44	55	86	55	54	69
		Total correct: 204			Total correct: **187**			Total correct: 182			Total correct: 180		
RF	BU	64	47	32	33	14	75	36	26	54	30	23	66
	TD	113	89	43	13	38	151	32	49	113	24	39	135
	COM	44	48	15	35	24	112	38	46	101	37	30	111
		Total correct: 168			Total correct: 183			Total correct: 186			Total correct: 180		
XGB	BU	79	51	13	48	30	44	49	28	39	37	43	39
	TD	111	120	14	39	70	93	32	73	89	43	68	87
	COM	41	55	11	56	56	59	52	61	72	46	53	79
		Total correct: **210**			Total correct: 177			Total correct: **194**			Total correct: **184**		

Table 8 (continued)

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			All levels		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
SVM	BU	52	67	24	54	27	41	40	17	59	59	22	38
	TD	107	92	46	70	46	86	67	70	57	80	44	74
	COM	44	43	20	72	45	54	67	60	58	76	29	73
		Total correct: 164			Total correct: 154			Total correct: 168			Total correct: 176		
NN	BU	48	50	45	34	57	31	30	64	22	42	43	34
	TD	84	97	64	44	92	66	62	95	37	52	71	75
	COM	46	28	33	51	72	48	45	85	55	55	60	63
		Total correct: 178			Total correct: 174			Total correct: 180			Total correct: 176		

Best solutions in each category are indicated in bold

The rows correspond to the true optimal HF methods while the columns to the predictions made by the various classification models in the test set (rolling origin evaluation)

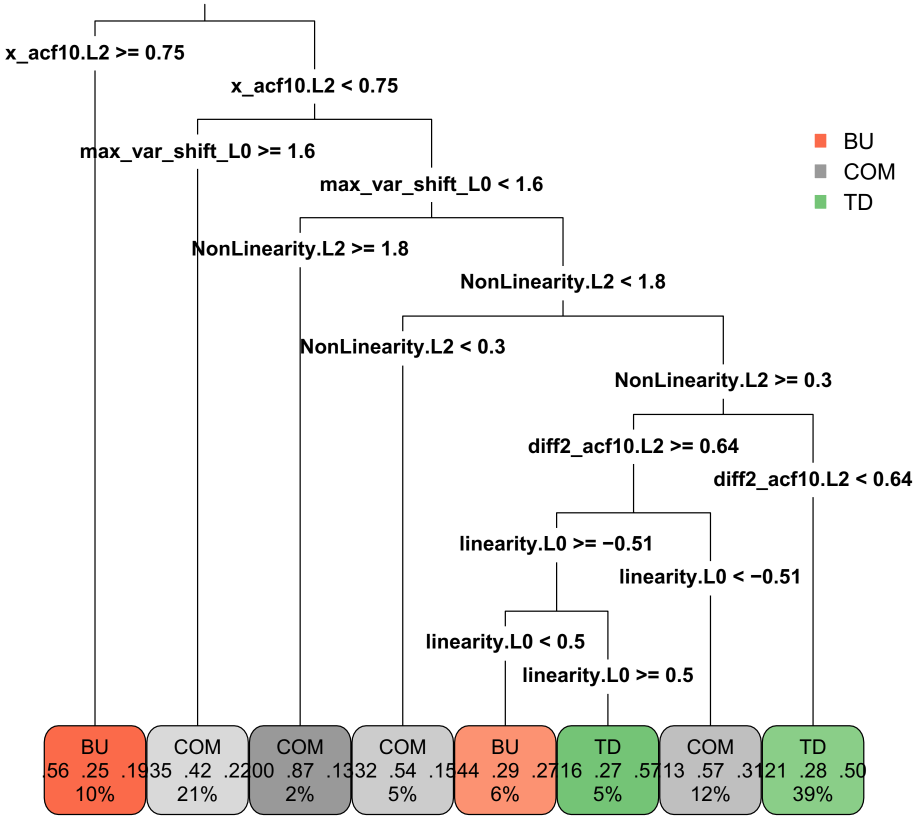

Fig. 12 Decision tree of the trained DT classification model across all series and hierarchies

16 periods for creating the labels of the classifiers in a rolling origin fashion considering one-step-ahead forecasts. We generate one-step-ahead forecasts to maximise the number of observations for training the ML classification models. Finally, the last 8 periods of data are used for evaluating the performance of the examined HF methods. Note that the set used for training the classifiers is highly imbalanced for most of the optimization criteria considered, with COM being the dominant HF method (COM is optimal at levels 1 and 2 in about 80% of the cases, while BU in the remaining 20%). The only exception is level 0 where BU and TD are the dominant HF approaches and similarly populated to each other (BU and TD are optimal at level 0 in about 40% of the cases each).

The results of the *Prison* data set are presented in Table 9, with their statistical significance being assessed in Fig. 13 using MCB tests. We should clarify that since the forecasts produced in this experiment refer to a single period, MASE and RMSSE results are identical. As seen, although the differences reported between the baseline HF methods and the CHF approaches are minor and, in most of the cases insignificant, the latter do manage to provide similar if not better ranks. Moreover, depending on the optimization criteria used, the CHF methods effectively select the most appropriate HF method per case, thus improving the forecasting accuracy accordingly. For instance, XGB-All displays the most accurate

Table 9 Forecasting performance of the HF methods examined in this study in terms of MASE/RMSSE for the *Prison* data set

HF Method	Level 0	Level 1	Level 2	Average
BU	0.187	0.509	0.672	0.456
TD	0.205	1.892	1.821	1.306
COM	0.201	**0.502**	**0.666**	0.456
LR-Avg	0.202	0.503	0.666	0.457
LDA-Avg	0.194	0.506	0.670	0.456
DT-Avg	0.201	**0.502**	**0.666**	0.456
RF-Avg	0.194	0.508	0.669	0.457
XGB-Avg	0.179	0.506	0.670	**0.452**
SVM-Avg	0.203	0.503	0.666	0.457
NN-Avg	0.201	**0.502**	**0.666**	0.456
LR-L0	0.215	1.202	1.221	0.879
LDA-L0	0.205	0.834	0.929	0.656
DT-L0	0.201	**0.502**	**0.666**	0.456
RF-L0	0.215	1.202	1.221	0.879
XGB-L0	**0.179**	1.586	1.524	1.096
SVM-L0	0.201	1.758	1.705	1.221
NN-L0	0.201	**0.502**	**0.666**	0.456
LR-L1	0.189	0.504	0.667	0.453
LDA-L1	0.188	0.506	0.670	0.455
DT-L1	0.201	**0.502**	**0.666**	0.456
RF-L1	0.191	0.505	0.668	0.455
XGB-L1	0.201	**0.502**	**0.666**	0.456
SVM-L1	0.201	**0.502**	**0.666**	0.456
NN-L1	0.201	**0.502**	**0.666**	0.456
LR-L2	0.187	0.506	0.668	0.454
LDA-L2	0.198	0.505	0.669	0.457
DT-L2	0.201	**0.502**	**0.666**	0.456
RF-L2	0.191	0.505	0.668	0.455
XGB-L2	0.201	**0.502**	**0.666**	0.456
SVM-L2	0.191	0.505	0.668	0.455
NN-L2	0.201	**0.502**	**0.666**	0.456

Best solutions in each category are indicated in bold

L0, L1, L2, and Avg indicate that the classification model used for implementing CHF is optimized with the objective of minimizing forecast error at level 0, level 1, level 2, and the average of all levels, respectively

results on average, XGB-L0 the best performance for level 0, DT-L1, XGB-L1, SVM-L1, NN-L1 the lowest forecast error for level 1, while DT-L2, XGB-L2, and NN-L2 the top accuracy for level 2. We also find that, similar to the *Sales* data set, XGB is generally more effective in completing the required classification tasks, outperforming on average the rest of the CHF methods. These results are justified by the confusion matrices of Table 10. We observe that XGB is the most precise classification model at level 0 and the average of all levels (75% accuracy), performing also similar to DT and NN at levels 1 and 2 (87.5%).

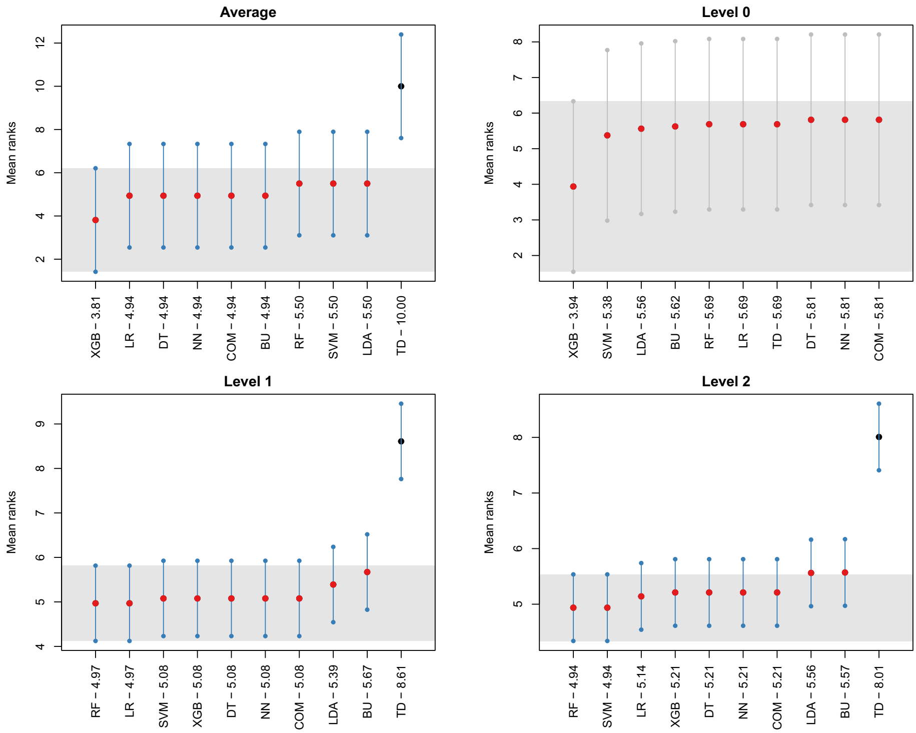

Fig. 13 MCB tests conducted on the HF methods examined in this study for the *Prison* data set. The results are presented for each optimization criterion separately, i.e., when the labels of the classification models are determined so that the forecasts produced are optimal in terms of level 0, level 1, level 2, or the average of all levels. In all cases, MASE/RMSSE is used for computing the ranks and a 95% confidence level is considered

Tourism data set

The *Tourism* data set is quarterly and spans over q1-1998 to q4-2017 (80 periods). We start producing base forecasts after the 6th year and use the following 36 periods for creating the labels of the classifiers in a rolling origin fashion considering two-step-ahead forecasts. We considered six years of data, i.e., 24 observations as the training set to provide enough observations to train the baseline forecasting model and considered two-step-ahead rolling forecasts to generate enough observations for the ML training data set. Note that two-step-ahead forecasts is often enough in practice for short term planning in the tourism industry. Moreover, since the selected horizon is larger than the one considered in the *Prison* data set, this setup allows us to further evaluate the performance of the CHF approach when longer forecast periods are considered. The last 20 periods of data are used for evaluating the performance of the examined HF methods. Similar to the *Prison* data set, the set used for training the classifiers is imbalanced, with COM being the dominant HF method for levels 1 and 2 (COM is optimal at levels 1 and 2 in about 75% of the cases, while BU in the remaining 25%), while TD for level 0 (TD is optimal at level 0 in about 60% of the cases, while BU and COM in 20% each).

Table 10 Confusion matrices of the classification models considered for forecasting the series of the *Prison* data set when the HF methods (labels) are determined so that the forecasts produced are optimal in terms of MASE/RMSSE at level 0, level 1, level 2, or the average of all levels

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			Average		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
LR	BU	2	2	0	1	0	0	0	0	1	1	0	3
	TD	2	2	0	0	0	0	0	0	0	0	0	0
	COM	0	0	0	1	0	6	1	0	6	1	0	3
		Total correct: 4			Total correct: **7**			Total correct: 6			Total correct: 4		
LDA	BU	1	1	2	1	0	0	0	0	1	1	0	3
	TD	1	1	2	0	0	0	0	0	0	0	0	0
	COM	0	0	0	4	0	3	4	0	3	2	0	2
		Total correct: 2			Total correct: 4			Total correct: 3			Total correct: 3		
DT	BU	0	0	4	0	0	1	0	0	1	0	0	4
	TD	0	0	4	0	0	0	0	0	0	0	0	0
	COM	0	0	0	0	0	7	0	0	7	0	0	4
		Total correct: 0			Total correct: **7**			Total correct: **7**			Total correct: 4		
RF	BU	2	2	0	1	0	0	1	0	0	2	0	2
	TD	2	2	0	0	0	0	0	0	0	0	0	0
	COM	0	0	0	2	0	5	2	0	5	3	0	1
		Total correct: 4			Total correct: 6			Total correct: 6			Total correct: 3		
XGB	BU	2	2	0	0	0	1	0	0	1	2	0	2
	TD	0	4	0	0	0	0	0	0	0	0	0	0
	COM	0	0	0	0	0	7	0	0	7	0	0	4
		Total correct: **6**			Total correct: **7**			Total correct: **7**			Total correct: **6**		

Table 10 (continued)

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			Average		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
SVM	BU	0	3	1	0	0	1	1	0	0	0	0	4
	TD	0	4	0	0	0	0	0	0	0	0	0	0
	COM	0	0	0	0	0	7	2	0	5	1	0	3
		Total correct: 4			Total correct: **7**			Total correct: 6			Total correct: 3		
NN	BU	0	0	4	0	0	1	0	0	1	0	0	4
	TD	0	0	4	0	0	0	0	0	0	0	0	0
	COM	0	0	0	0	0	7	0	0	7	0	0	4
		Total correct: 0			Total correct: **7**			Total correct: **7**			Total correct: 4		

Best solutions in each category are indicated in bold

The rows correspond to the true optimal HF methods while the columns to the predictions made by the various classification models in the test set (rolling origin evaluation)

Table 11 Forecasting performance of the HF methods examined in this study in terms of MASE for the *Tourism* data set

HF Method	Level 0	Level 1	Level 2	Average
BU	1.610	1.224	0.963	1.266
TD	**0.923**	1.586	1.280	1.263
COM	1.302	**1.099**	**0.937**	**1.113**
LR-Avg	1.347	1.288	1.039	1.225
LDA-Avg	1.181	1.220	1.026	1.142
DT-Avg	1.302	**1.099**	**0.937**	**1.113**
RF-Avg	1.302	**1.099**	**0.937**	**1.113**
XGB-Avg	1.302	**1.099**	**0.937**	**1.113**
SVM-Avg	1.098	1.341	1.102	1.180
NN-Avg	1.302	**1.099**	**0.937**	**1.113**
LR-L0	0.947	1.528	1.238	1.238
LDA-L0	1.355	1.224	1.016	1.198
DT-L0	1.302	**1.099**	**0.937**	**1.113**
RF-L0	**0.923**	1.586	1.280	1.263
XGB-L0	**0.923**	1.586	1.280	1.263
SVM-L0	1.077	1.461	1.207	1.248
NN-L0	1.302	**1.099**	**0.937**	**1.113**
LR-L1	1.356	1.121	0.943	1.140
LDA-L1	1.445	1.149	0.948	1.181
DT-L1	1.302	**1.099**	**0.937**	**1.113**
RF-L1	1.302	**1.099**	**0.937**	**1.113**
XGB-L1	1.302	**1.099**	**0.937**	**1.113**
SVM-L1	1.370	1.123	0.943	1.145
NN-L1	1.302	**1.099**	**0.937**	**1.113**
LR-L2	1.351	1.115	0.939	1.135
LDA-L2	1.423	1.144	0.946	1.171
DT-L2	1.302	**1.099**	**0.937**	**1.113**
RF-L2	1.320	1.101	0.937	1.119
XGB-L2	1.302	**1.099**	**0.937**	**1.113**
SVM-L2	1.320	1.101	0.937	1.119
NN-L2	1.302	**1.099**	**0.937**	**1.113**

Best solutions in each category are indicated in bold

The results of the *Tourism* data set are presented in Tables 11 and 12, with their statistical significance being assessed in Figs. 14 and 15, respectively. As seen, the relative ranks of the HF methods examined are in agreement according to both accuracy measures used. Moreover, CHF approaches always result in significantly better average ranks than BU, being also better than TD at levels 2, 1, and the average of all levels. Therefore, we conclude that even when the differences between the proposed HF method and the baseline ones are small, CHF manages to effectively select the reconciliation method that best matches the needs of the examined forecasting task. These results can be verified by the exceptional classification performance of the classifiers considered, summarized in Tables 13 and 14. We observe that XGB and RF report an accuracy of 90% or higher across all the hierarchical levels examined, with the rest of the models achieving also comparable results in most of the cases.

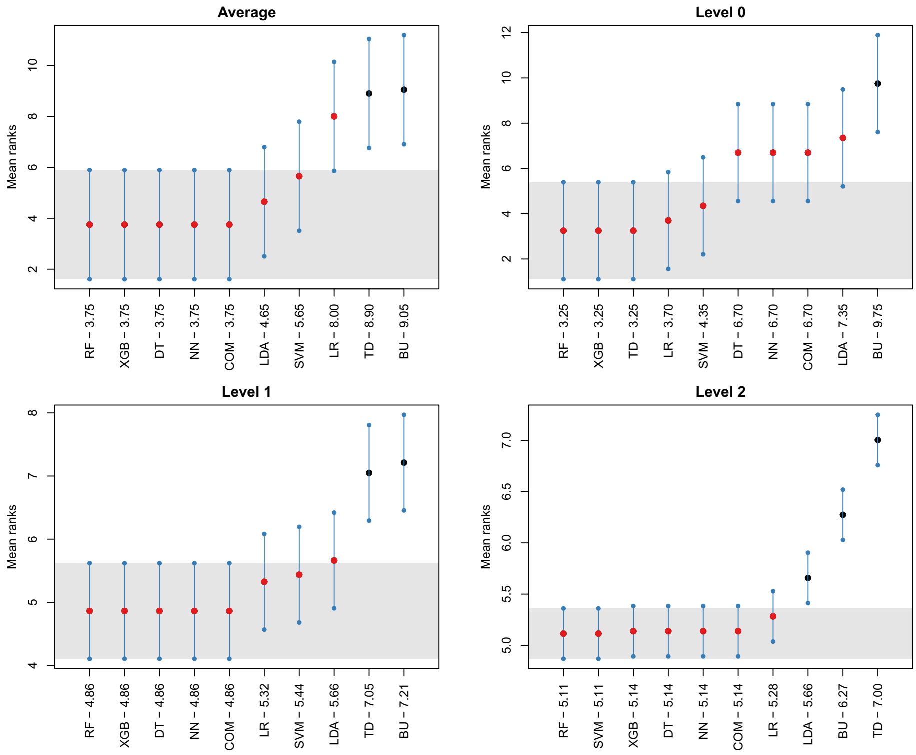

Fig. 14 MCB tests conducted on the HF methods examined in this study for the *Tourism* data set. The results are presented for each optimization criterion separately, i.e., when the labels of the classification models are determined so that the forecasts produced are optimal in terms of level 0, level 1, level 2, or the average of all levels. In all cases, MASE is used for computing the ranks and a 95% confidence level is considered

Appendix E: Computational time

In order to evaluate the performance of CHF in terms of computational requirements, we recorded the computational times required to (i) execute the HF methods considered in our study for reconciling the base forecasts, (ii) execute the classification models utilized for determining the optimal HF method, and (iii) estimate the time series features exploited by the classifiers. Table 15 summarizes our findings for the case of the *Sales* data set as it is the largest in size from the examined ones (55 hierarchies × 9 rolling origin evaluations = 495 instances) and can therefore provide more representative results. The times reported refer to the seconds elapsed on average for forecasting a complete hierarchy and were computed using a system of the following characteristics: MacBook Pro, 2.7 GHz Dual-Core Intel Core i5 processor, and 8 GB 1867 MHz DDR3 Memory RAM.

As seen, COM is the most computationally intensive HF method (6.35 s), followed by BU (5.17 s). On the other hand, as expected, TD is much faster, requiring less than 0.2 s to run. As a result, a traditional forecasting approach that does not exploit model selection and requires computing all three HF methods before identifying the best possible, would take a total of 11.68 ss to execute, on average. Regarding the classifiers, XGB, the most

Table 12 Forecasting performance of the HF methods examined in this study in terms of RMSSE for the *Tourism* data set

HF Method	Level 0	Level 1	Level 2	Average
BU	1.687	1.345	1.066	1.366
TD	**1.017**	1.78	1.427	1.408
COM	1.389	**1.219**	**1.036**	**1.215**
LR-Avg	1.33	1.371	1.123	1.275
LDA-Avg	1.558	1.297	1.054	1.303
DT-Avg	1.389	**1.219**	**1.036**	**1.215**
RF-Avg	1.389	**1.219**	**1.036**	**1.215**
XGB-Avg	1.389	**1.219**	**1.036**	**1.215**
SVM-Avg	1.414	1.221	1.036	1.224
NN-Avg	1.389	**1.219**	**1.036**	**1.215**
LR-L0	1.065	1.741	1.392	1.399
LDA-L0	1.176	1.682	1.362	1.407
DT-L0	1.389	**1.219**	**1.036**	**1.215**
RF-L0	**1.017**	1.780	1.427	1.408
XGB-L0	**1.017**	1.780	1.427	1.408
SVM-L0	1.090	1.712	1.393	1.398
NN-L0	1.389	**1.219**	**1.036**	**1.215**
LR-L1	1.442	1.243	1.041	1.242
LDA-L1	1.425	1.24	1.041	1.235
DT-L1	1.389	**1.219**	**1.036**	**1.215**
RF-L1	1.389	**1.219**	**1.036**	**1.215**
XGB-L1	1.389	**1.219**	**1.036**	**1.215**
SVM-L1	1.467	1.245	1.041	1.251
NN-L1	1.389	**1.219**	**1.036**	**1.215**
LR-L2	1.435	1.236	1.040	1.237
LDA-L2	1.506	1.262	1.047	1.272
DT-L2	1.389	**1.219**	**1.036**	**1.215**
RF-L2	1.406	1.222	1.037	1.222
XGB-L2	1.389	**1.219**	**1.036**	**1.215**
SVM-L2	1.406	1.222	1.037	1.222
NN-L2	1.389	**1.219**	**1.036**	**1.215**

Best solutions in each category are indicated in bold

accurate classification model of our study, is executed in 0.25 seconds, while simpler classification models, such as LR and LDA, are executed even faster (0.11 and 0.06 s, respectively). Given that the computation of the time series features requires 4.13 s, the complete implementation of CHF takes on average between 4.35 and 10.82 s, depending on the complexity of the classification model used and the HF method selected. Thus, the proposed method can effectively reduce computational requirements from 7 to 63% compared to an approach where all three HF methods are used to reconcile the base forecasts, while at the same time significantly improving the overall forecasting accuracy.

We should highlight that the computational cost of CHF is mostly driven by the time required for producing the base forecasts and implementing the HF method of choice (computing the features and executing the classifier accounts for less than 40% of the computational time on average). As a result, the computational savings of the proposed approach

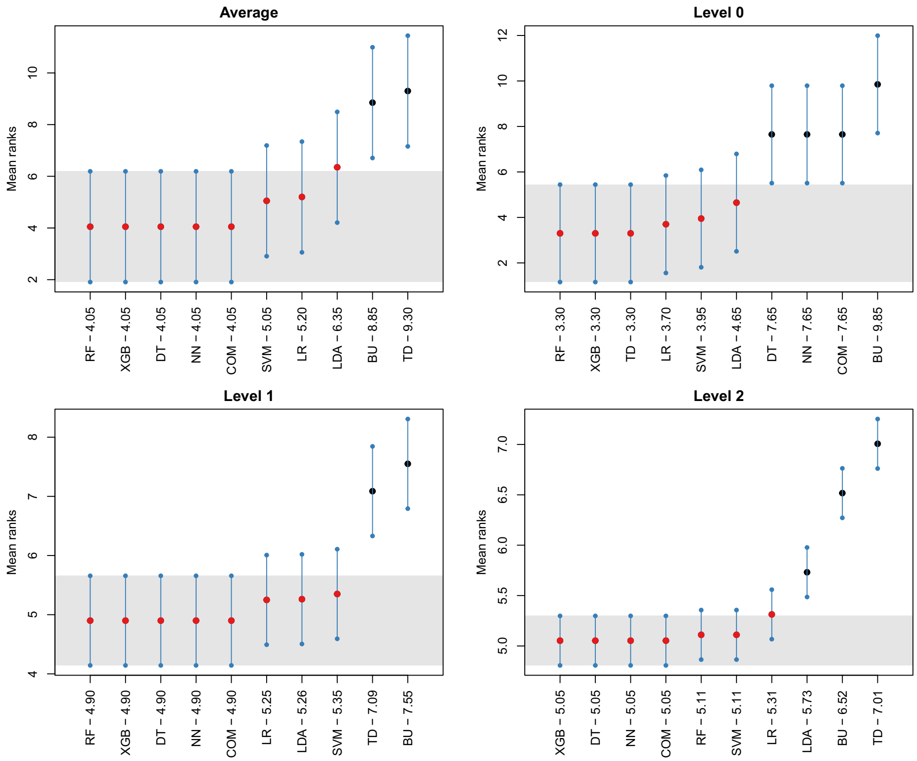

Fig. 15 MCB tests conducted on the HF methods examined in this study for the *Tourism* data set. The results are presented for each optimization criterion separately, i.e., when the labels of the classification models are determined so that the forecasts produced are optimal in terms of level 0, level 1, level 2, or the average of all levels. In all cases, RMSSE is used for computing the ranks and a 95% confidence level is considered

can grow further when more sophisticated methods are used for producing the base forecasts or when hierarchies that consist of more levels or series are being forecast.

Appendix F: Benchmarking CHF against similar approaches

In order to evaluate the performance of CHF against other approaches that employ classifications models to determine the optimal HF method, we implemented the approach described in Nenova and May (2016) as, according to Table 1, it is the closest work in the literature to our study. Nenova and May (2016) (from now on called N & M) conducted an experiment on 104 two-level hierarchical series for various products. They used rank predictors, the correlation of the series of the bottom level, and a combination of them as input variables to select the most appropriate HF method from BU and three variations of TD, as implemented in the *hts* package for Hyndman et al. (2020b). They did so in two different paradigms, i.e., by using a four label classification model to directly select the most appropriate HF method, and by using a two-stage approach where they first determined whether BU or TD should be used and then selected the best TD variant, if needed.

Table 13 Confusion matrices of the classification models considered for forecasting the series of the *Tourism* data set when the HF methods (labels) are determined so that the forecasts produced are optimal in terms of MASE at level 0, level 1, level 2, or the average of all levels

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			Average		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
LR	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	8	1	0	0	0	0	0	0	0	0	0
	COM	0	1	0	2	0	7	2	0	7	7	2	1
		Total correct: 8			Total correct: 7			Total correct: 7			Total correct: 1		
LDA	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	3	2	4	0	0	0	0	0	0	0	0	0
	COM	1	0	0	5	0	4	4	0	5	0	2	8
		Total correct: 2			Total correct: 4			Total correct: 5			Total correct: 8		
DT	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	0	9	0	0	0	0	0	0	0	0	0
	COM	0	0	1	0	0	9	0	0	9	0	0	10
		Total correct: 1			Total correct: **9**			Total correct: **9**			Total correct: **10**		
RF	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	9	0	0	0	0	0	0	0	0	0	0
	COM	0	1	0	0	0	9	1	0	8	0	0	10
		Total correct: **9**			Total correct: **9**			Total correct: 8			Total correct: **10**		
XGB	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	9	0	0	0	0	0	0	0	0	0	0
	COM	0	1	0	0	0	9	0	0	9	0	0	10
		Total correct: **9**			Total correct: **9**			Total correct: **9**			Total correct: **10**		

Table 13 (continued)

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			Average		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
SVM	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	1	7	1	0	0	0	0	0	0	0	0	0
	COM	0	1	0	3	0	6	1	0	8	0	4	6
		Total correct: 7			Total correct: 6			Total correct: 8			Total correct: 6		
NN	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	0	9	0	0	0	0	0	0	0	0	0
	COM	0	0	1	0	0	9	0	0	9	0	0	10
		Total correct: 1			Total correct: **9**			Total correct: **9**			Total correct:**10**		

Best solutions in each category are indicated in bold

The rows correspond to the true optimal HF methods while the columns to the predictions made by the various classification models in the test set (rolling origin evaluation)

Table 14 Confusion matrices of the classification models considered for forecasting the series of the *Tourism* data set when the HF methods (labels) are determined so that the forecasts produced are optimal in terms of RMSSE at level 0, level 1, level 2, or the average of all levels

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			Average		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
LR	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	9	1	0	0	0	0	0	0	0	0	0
	COM	0	0	0	2	0	7	2	0	7	0	2	8
		Total correct: 9			Total correct: 7			Total correct: 7			Total correct: 8		
LDA	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	2	8	0	0	0	0	0	0	0	0	0	0
	COM	0	0	0	1	0	8	4	0	5	5	0	5
		Total correct: 8			Total correct: 8			Total correct: 5			Total correct: 5		
DT	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	0	10	0	0	0	0	0	0	0	0	0
	COM	0	0	0	0	0	9	0	0	9	0	0	10
		Total correct: 0			Total correct: **9**			Total correct: **9**			Total correct: **10**		
RF	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	10	0	0	0	0	0	0	0	0	0	0
	COM	0	0	0	0	0	9	1	0	8	0	0	10
		Total correct: **10**			Total correct: **9**			Total correct: 8			Total correct: **10**		
XGB	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	10	0	0	0	0	0	0	0	0	0	0
	COM	0	0	0	0	0	9	0	0	9	0	0	10
		Total correct: **10**			Total correct: **9**			Total correct: **9**			Total correct: **10**		

Table 14 (continued)

Models	Actual	Prediction											
		Level 0			Level 1			Level 2			Average		
		BU	TD	COM	BU	TD	COM	BU	TD	COM	BU	TD	COM
SVM	BU	0	0	0	1	0	0	0	0	1	0	0	0
	TD	1	9	0	0	0	0	0	0	0	0	0	0
	COM	0	0	0	3	0	6	1	0	8	2	0	8
		Total correct: 9			Total correct: 7			Total correct: 8			Total correct: 8		
NN	BU	0	0	0	0	0	1	0	0	1	0	0	0
	TD	0	0	10	0	0	0	0	0	0	0	0	0
	COM	0	0	0	0	0	9	0	0	9	0	0	10
		Total correct: 0			Total correct: **9**			Total correctTotal correct: **9**			Total correct: **10**		

Best solutions in each category are indicated in bold

The rows correspond to the true optimal HF methods while the columns to the predictions made by the various classification models in the test set (rolling origin evaluation)

Table 15 Computational time (seconds) required to run the experiment for the *Sales* data set

Task	BU	TD	COM	LR	LDA	DT	RF	XGB	SVM	NN	Features
Time	5.17	0.16	6.35	0.11	0.06	0.18	0.34	0.25	0.23	0.19	4.13

Table 16 Forecasting performance of the Nenova and May (2016) (N & M) approach for the *Sales* data set over the method proposed in the present study (CHF).

Methods	Level 0	Level 1	Level 2	Average
CHF				
LR	0.508	0.858	0.869	0.844
LDA	0.509	0.855	0.867	0.841
DT	0.495	0.855	0.866	0.840
RF	0.470	0.846	0.848	0.721
XGB	0.466	**0.820**	**0.828**	**0.705**
SVM	**0.457**	0.850	0.850	0.719
NN	0.495	0.855	0.866	0.840
N & M				
LR	0.527	1.074	1.044	1.013
LDA	0.525	1.076	1.048	1.016
DT	0.531	1.039	1.035	1.001
RF	0.545	1.045	1.039	1.006
XGB	0.512	1.027	1.014	0.982
SVM	0.546	1.065	1.054	1.021
NN	0.531	1.039	1.035	1.001

Best solutions in each category are indicated in bold

Accuracy is measured using MASE. The classification models employed in N & M and CHF are optimized so that the forecasts are optimal for the average of the three hierarchical levels

They used LR, LDA, DT, SVM, and C5.0 classification models to select the optimal HF method. They reported the results only for the models where they used time series correlations as predictors and found that LDA provided the most accurate forecasts. We replicated the N & M approach and implemented it on the *Sales* data set since it is the primary data set considered in the present study. We used correlation predictors at the bottom level and included their expectation, minimum value, maximum value, and variance as additional leading indicators. We only implemented the direct, multi-label classification approach of N & M since, in our study, we use TD, BU, and COM as the potential HF methods, meaning that the two-stage classification approach of N & M is not applicable in our settings. Although N & M did not employ the COM method, we believe it is important to include it in the evaluation as the results of several empirical studies have shown that COM is often a top-performing HF method in various settings (Hyndman et al., 2011; Abolghasemi et al., 2019; Spiliotis et al., 2020). Moreover, by doing so, we can directly compare the results of the CHF approach to that of N & M by using time series correlations as predictors rather than time series characteristics. Table 16 summarizes the performance of the N & M approach when employed for the *Sales* data set. Our results indicate that CHF outperforms the N & M method across all hierarchical levels regardless of the classification technique used. This finding suggests that time series features may be more useful than time series correlations for identifying the most suitable HF method.

Author Contributions MA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing, review. RJH: Conceptualization, Investigation, Methodology, Project administration, Supervision, Validation, Writing, review. ES: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing, review. CB: Conceptualization, Investigation, Methodology, Supervision, Validation, Writing, review

Funding Not applicable.

Data availability The Tourism and Prison hierarchical data sets are publicly available. The sales data is not publicly available for confidentiality reasons.

Code availability The R code is available at https://github.com/mahdiabolghasemi/Conditional-reconciliation-in-HF

Declarations

Conflict of interest The Authors declare that there is no conflict of interest.

Consent to publication All authors participated in this study give the publisher the permission to publish this work.

Ethical approval Waivers, Not applicable.

References

Abolghasemi, M., Hyndman, R. J., Tarr, G., & Bergmeir, C. (2019). Machine learning applications in time series hierarchical forecasting. arXiv preprint arXiv:1912.00370

Abolghasemi, M., Beh, E., Tarr, G., & Gerlach, R. (2020). Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion. *Computers & Industrial Engineering, 142,* 106380.

Abolghasemi, M., Hurley, J., Eshragh, A., & Fahimnia, B. (2020). Demand forecasting in the presence of systematic events: Cases in capturing sales promotions. *International Journal of Production Economics, 230,* 107892.

Abouarghoub, W., Nomikos, N. K., & Petropoulos, F. (2018). On reconciling macro and micro energy transport forecasts for strategic decision making in the tanker industry. *Transportation Research Part E: Logistics and Transportation Review, 113,* 225–238.

Adya, Monica, Armstrong, J Scott, Collopy, Fred, & Kennedy, Miles. (2000). An application of rule-based forecasting to a situation lacking domain knowledge. *International Journal of Forecasting, 16*(4), 477–484.

Adya, M., Collopy, F., Armstrong, J. S., & Kennedy, M. (2001). Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting, 17*(2), 143–157.

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting, 25*(1), 146–166.

Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R. J., & Affan, M. (2020). Hierarchical Forecasting. In F. Peter (Ed.), *Macroeconomic forecasting in the era of big data: Theory and practice* (pp. 689–719). Springer.

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting, 36*(1), 197–200.

Burba, D., & Chen, T. (2021). A trainable reconciliation method for hierarchical time-series. arXiv preprint arXiv:2101.01329

Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications, 112,* 353–371.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). ACM.

Chen, H., & Boylan, J. E. (2009). The effect of correlation between demands on hierarchical forecasting. *Advances in business and management forecasting* (pp. 173–188). Emerald Group Publishing Limited.

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.

Dangerfield, B. J., & Morris, J. S. (1992). Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting, 8*(2), 233–241.

Demolli, H., Sakir Dokuz, A., Ecemis, A., & Gokcek, M. (2019). Wind power forecasting based on daily wind speed data using machine learning algorithms. *Energy Conversion and Management, 198,* 111823.

Fildes, R. (2001). Beyond forecasting competitions. *International Journal of Forecasting, 17*(4), 556–560.

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

Fliedner, G. (1999). An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers & Operations Research, 26*(10–11), 1133–1149.

Fliedner, G. (2001). Hierarchical forecasting: Issues and use guidelines. *Industrial Management & Data Systems, 101*(1), 5–12.

Fulcher, B. D., Little, M. A., & Jones, N. S. (2013). Highly comparative time-series analysis: The empirical structure of time series and their methods. *Journal of the Royal Society Interface, 10*(83), 20130048.

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

Garland, J., James, R., & Bradley, E. (2014). Model-free quantification of time-series predictability. *Physical Review E, 90*(5), 052910.

Ghobbar, A. A., & Friend, C. H. (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: A predictive model. *Computers & Operations Research, 30*(14), 2097–2114.

Giir Ali, O., Sayin, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications, 36*(10), 12340–12348.

Goerg, G. (2013). Forecastable component analysis. In *International conference on machine learning* (pp. 64-72).

Gross, C. W., & Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting, 9*(3), 233–254.

Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. In *Advances in neural information processing systems* (pp. 507–513).

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice, 3rd edn. OTexts. http://OTexts.com/fpp3

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2020a). forecast: Forecasting functions for time series and linear models. R package version *8*(12). http://pkg.robjhyndman.com/forecast

Hyndman, R., Kang, Y., Montero-Manso, P., Talagala, T., Wang, E., & Yang, Y. (2019). Tsfeatures: Time series feature extraction. In *R package version 1.0.1*. https://pkg.robjhyndman.com/tsfeatures/

Hyndman, R., Lee, A., Wang, E., & Wickramasuriya, S. (2020b). hts: Hierarchical and grouped time series. R package version 6.0.0. https://CRAN.R-project.org/package=hts

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis, 55*(9), 2579–2589.

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.

Hyndman, R. J., Lee, A. J., & Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis, 97,* 16–32.

Jeon, J., Panagiotelis, A., & Petropoulos, F. (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research, 279*(2), 364–379.

Kahn, K. B. (1998). Revisiting top-down versus bottom-up forecasting. *The Journal of Business Forecasting, 17*(2), 14.

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.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 package for kernel methods in R. *Journal of Statistical Software, 11*(9), 1–20.

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting, 32*(3), 788–803.

Koning, A. J., Hans Franses, P., Hibon, M., & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting, 21*(3), 397–409.

Kourentzes, N., & Athanasopoulos, G. (2019). Cross-temporal coherent forecasts for Australian tourism. *Annals of Tourism Research, 75,* 393–409.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics, 54*(1–3), 159–178.

Lemke, C., & Gabrys, B. (2010). Meta-learning for time series forecasting and forecast combination. *Neurocomputing, 73*(10–12), 2006–2016.

Liaw, A., & Wiener, M. (2002). Classification and regression by ran-domForest. *R News, 2*(3), 18–22.

Liu, X., Jiang, A., Xu, N., & Xue, J. (2016). Increment entropy as a measure of complexity for time series. *Entropy, 18*(1), 22.

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

Mahajan, V., & Wind, Y. (1988). New product forecasting models: Directions for research and implementation. *International Journal of Forecasting, 4*(3), 341–358.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M5 competition: Background, organization and implementation. Working paper.

Makridakis, S., Hyndman, R. J., & Petropoulos, F. (2020). Forecasting in social settings: The state of the art. *International Journal of Forecasting, 36*(1), 15–28.

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.

Mancuso, P., Piccialli, V., & Sudoso, A. M. (2020). A machine learning approach for forecasting hierarchical time series. arXiv preprint arXiv:2006.00630

Meade, N. (2000). Evidence for the selection of forecasting methods. *Journal of Forecasting, 19*(6), 515–535.

Mircetic, D., Rostami-Tabar, B., Nikolicic, S., & Maslaric, M. (2021). Forecasting hierarchical time series in supply chains: An empirical investigation. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2021.1896817.

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.

Nenova, Z. D., & May, J. H. (2016). Determining an optimal hierarchical forecasting model based on the characteristics of the data set. *Journal of Operations Management, 44,* 62–68.

Nielsen, D. (2016). Tree boosting with xgboost-why does xgboost win every machine learning competition?. MA thesis. NTNU.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research, 268*(2), 545–554.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., & Nikolopoulos, K. (2014). Horses for courses' in demand forecasting. *European Journal of Operational Research, 237*(1), 152–163.

Pooya, A., Pakdaman, M., & Tadj, L. (2019). Exact and approximate solution for optimal inventory control of two-stock with reworking and forecasting of demand. *Operational Research, 19*(2), 333–346.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyper-parameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(3), e1301.

Reid, D. J. (1972). A comparison of forecasting techniques on economic time series. Forecasting in action. Operational Research Society and the Society for Long Range Planning.

Schafer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology *4*(1).

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting, 36*(1), 75–85.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.

Spiliotis, E., Abolghasemi, M., Hyndman, R. J., Petropou-los, F., & Assimakopoulos, V. (2020). Hierarchical forecast reconciliation with machine learning. arXiv preprint arXiv:2006.02043

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.

Spiliotis, E., Petropoulos, F., & Assimakopoulos, V. (2019). Improving the forecasting performance of temporal hierarchies. *PloS ONE, 14*(10), e0223422.

Spiliotis, E., Petropoulos, F., Kourentzes, N., & Assimakopoulos, V. (2020). Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption. *Applied Energy, 261,* 114339.

Syntetos, A. A., Nikolopoulos, K., & Boylan, J. E. (2010). Judging the judges through accuracy-implication metrics: The case of inven-tory forecasting. *International Journal of Forecasting, 26*(1), 134–143.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450.

Wang, X., Smith, K., & Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery, 13*(3), 335–364.

Wang, X., Smith-Miles, K., & Hyndman, R. (2009). Rule induction for forecasting method selection: Meta-learning the characteristics of uni-variate time series. *Neurocomputing, 72*(10–12), 2581–2594.

Wickramasuriya, S. L., Athanasopoulos, G., & Hynd-man, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *J American Statistical Association, 114*(526), 804–819.

Widiarta, H., Viswanathan, S., & Piplani, R. (2007). On the effectiveness of top-down strategy for forecasting autoregressive demands. *Naval Research Logistics (NRL), 54*(2), 176–188.

Widiarta, H., Viswanathan, S., & Piplani, R. (2008). Forecasting item-level demands: An analytical evaluation of top-down versus bottom-up forecasting in a production-planning framework. *IMA Journal of Management Mathematics, 19*(2), 207–218.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.