A Natural Gas Consumption Forecasting System for Continual Learning Scenarios based on Hoeffding Trees with Change Point Detection Mechanism

Radek Svoboda[a,*], Sebastián Basterrech[b], Jędrzej Kozal[c], Jan Platoš[a], Michał Woźniak[c]

[a]*Faculty of Electrical Engineering and Computer Science, VŠB–Technical University of Ostrava, 17. listopadu 2172/15, Ostrava, 70833, Czechia*
[b]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark*
[c]*Department of Systems and Computer Networks, University of Science and Technology, Wroclaw, Poland*

Abstract

Forecasting natural gas consumption, considering seasonality and trends, is crucial in planning its supply and consumption and optimizing the cost of obtaining it, mainly by industrial entities. However, in times of threats to its supply, it is also a critical element that guarantees the supply of this raw material to meet individual consumers' needs, ensuring society's energy security. This article introduces a novel multistep forecasting of natural gas consumption with change point detection integration for model collection selection with continual learning capabilities using data stream processing. The performance of the forecasting models based on the proposed approach is evaluated in a complex real-world use case of natural gas consumption

---
[*]Corresponding author

Email addresses: `radek.svoboda@vsb.cz` (Radek Svoboda), `sebbas@dtu.dk` (Sebastián Basterrech), `jedrzej.kozal@pwr.edu.pl` (Jędrzej Kozal), `jan.platos@vsb.cz` (Jan Platoš), `michal.wozniak@pwr.edu.pl` (Michał Woźniak)

forecasting, furthermore the methodology generability was also verified in an electricity load forecasting task. We employed Hoeffding tree predictors as forecasting models and the Pruned Exact Linear Time (PELT) algorithm for the change point detection procedure. The change point detection integration enables selecting a different model collection for successive time frames. Thus, three model collection selection procedures are defined and evaluated for forecasting scenarios with various densities of detected change points. These models were compared with change point agnostic baseline approaches and also deep learning models. Our experiments show that the proposed approach provides superior results compared to deep learning models for both datasets and also that fewer change points result in a lower forecasting error regardless of the model collection selection procedure employed.

Keywords: incremental learning, time series forecasting, multivariate time series, change point detection, machine learning, data stream processing

1. Introduction

As a result of ongoing trends, the energy industry is experiencing a growing demand for consumption forecasting tools. In the past, these tools focused mainly on electricity load forecasting, as they presented complex and costly storage challenges. Consequently, the development of similar tools for other areas was postponed. However, the importance of natural gas is now on the rise, driven by the increasing environmental initiatives undertaken by governments and private companies. Despite this, effective logistics and delivery planning in the natural gas sector presents significant challenges.

A primary driver for energy consumption forecasting stems from the

prevalent use of take-or-pay contracts and clauses in commodity trading [1, 2, 3]. These contractual agreements require customers to pay for the specified quantity, regardless of actual usage. The financial consequences of such agreements underscore the importance of improving the quality of the forecasting models. Clearly, the vital landscape of energy consumption shows a need for more effective forecasting models with continual adaptation capabilities. This adaptability ensures that the models remain relevant and effective in capturing the nuances of changing consumption patterns, market dynamics, and external influences. This responsiveness to new data improves the overall reliability of the system. Creating a more resilient framework capable of addressing the challenges inherent in the energy sector. Therefore, recognizing the pivotal role of continual model adaptation is crucial for stakeholders navigating the intricacies of energy consumption forecasting in the context of take-or-pay contracts.

Change point detection techniques are useful for understanding and predicting patterns in time series forecasting. A change point event refers to a specific moment in a sequence where the underlying behavior or trend shifts abruptly. Various factors can cause these shifts, such as policy changes, economic events, technological advancements, or changes in consumer behavior. Identifying change points is crucial to developing an accurate and robust forecasting model. Change points can significantly impact the underlying distribution of time series. In the particular case of energy consumption, it can cause modifications in the near-term energy demand. Furthermore, understanding change points and seasonality in time series forecasting enables energy industry stakeholders to make more informed decisions. By

identifying abrupt changes in consumption behavior, energy providers can anticipate shifts in energy demand, adjust their strategies accordingly, and improve energy grid management.

Recently, Continual Learning (CL) holds great importance in the field of time series forecasting. With technological advances and the availability of large-scale datasets, CL allows forecasting methods to incorporate advanced approaches and continuously integrate new information, enhancing their forecasting capabilities. The ability to adapt and learn from new data and methodologies ensures that forecasting tools remain reliable and effective in meeting the evolving demands of the energy industry. By continually updating and refining predictive models in the context of CL, predictive tools can capture distribution changes in input patterns and provide different weights to the latest data, enabling more accurate predictions of energy demand.

The main contributions of the work are listed below.

(i) We introduce a general framework for multistep forecasting of multivariate time series as CL task. The innovative framework is composed of several hierarchical independent modules, such as dataset preprocessing, change point detection, model collection setup, model collection selection schema, and model training, which are fully described in the paper. The modular architecture enhances flexibility and makes it easy to exchange specific modules if the characteristics of the problem change.

(ii) We study the performance of the framework in a wide range of scenarios. We investigate several approaches for forecasting natural gas

4

consumption and electricity load, using the integration of the detection of change points with the selection approach of model collection (with or without an error feedback loop).

(iii) We present a sensitive analysis of the integration of the PELT algorithm into the CL system. We also discuss the threat to validity of our research. The experimental analysis includes a comparison with two baseline approaches that do not consider change point detection and also with deep learning forecasting models. Furthermore, we study the impact of number of detected change points on the overall predictive performance. The evaluation was carried out on recently collected real-world dataset from the energy distribution domain. We evaluated the developed system using three standard metrics (MSE, MAE and SMAPE) and Diebold-Mariano test to better capture the time series behavior.

2. Related works

In this section, we focus on four pivotal topics central to our research. Our primary emphasis lies in forecasting natural gas consumption, given its vital role in energy planning and trading. Additionally, we delve into the literature of the CL paradigm in the area of energy consumption forecasting. The Hoeffding Decision Tree is explored for its algorithmic relevance, and we review the literature of Change Point Detection methods for identifying abrupt shifts in the time series data. By synthesizing insights from these areas, our study aims to build a solid foundation for improving the predictive capabilities of natural gas consumption models.

5

2.1. Natural gas consumption forecasting

Statistical methods have gained popularity in forecasting gas consumption during the late 90s [4]. A nonlinear regression model was applied to model the gas consumption behavior of individual and small commercial customers [5]. Due to the success of Neural Networks (NNs) for solving machine learning problems, several works also applied their potential. Oil market data and weather information were studied to create a forecasting model employing NNs for Belgian gas consumption [6]. Other families of NNs have also been studied, a two-stage model that combined feedforward and functional link NNs [7], while a hybrid of traditional networks and fuzzy NNs was developed [8]. Hourly sampled data was analyzed using linear methods, support vector machines, and forward NNs [9]. The authors showed how different time windows of the day present different consumption trends. Another learning application was over the gas consumption in Turkish regions [10, 11]. The SARIMAX model, multilayer perceptron ANN, and RBF models were evaluated for the prediction of daily gas consumption [10]. The impact of including exogenous variables in specific forecasting models for the gas consumption was also investigated, especially the relevance of the air temperature feature [10].

Multilayer feedforward neural network also showed to be effective as forecasting method of hourly gas consumption when the input layer has included features with weather data information [12]. However, the relationship between external variables and gas consumption was not always evident and direct. Several studies considered the thermal memory capacity of the buildings. For example, air temperature and solar radiation may have some delays

6

in the impacts of gas consumption [9, 13]. Forecasting based on the LSTM model and evolutionary algorithms for hyperparameter optimization were also studied [14]. Although NNs have been widely applied to solve the forecasting problem, other tools have also reported good results. Support vector machines were also employed with structure calibration [15]. According to their results, kernel-based methods outperformed classic backpropagation models in predicting daily gas consumption. Two reviews of the literature on the area of natural gas forecasting are available in [16, 17].

2.2. Continual Learning

Despite recent advances in the machine learning domain and the large number of applications, most learning tools are designed to model stationary learning data. The CL paradigm has been introduced to overcome some classic machine learning limitations. A CL approach defines an algorithm that can learn in incremental mode, which is suitable for not stationary data modeling [18]. The greatest challenge in such a setting is overcoming catastrophic forgetting [19] – a tendency for NNs to quickly forget what they have learned from previously seen data, especially when those data are no longer available. The optimal solution for the CL problem was shown to be NP-hard [20], and many heuristics have been proposed so far. Most of the current CL methods are based on rehearsal [21]. In such methods, historical samples are stored in a small buffer and could be used to prevent forgetting when training with new data. Rehearsal algorithms can be quickly adapted to the online setting by utilization of reservoir sampling [22]. In [23], instead of randomly selecting the learning examples from the rehearsal buffer when learning a new task, samples were selected based on the loss change

7

when performing network updates. The authors of [24] observed a shift of old data representations in the network feature space during training with new learning examples. They also introduced a method to correct this problem. The work of [25] provided theoretical insight into the risk of overfitting when using a small memory buffer. Dark experience replay [26] extended the standard rehearsal to save not only historical samples, but also logits with predictions that are later used for knowledge distillation. This method was later improved [27] by introducing bias correction, adding pre-training for future tasks, and updating existing logits in memory. Other approaches to CL can include the inclusion of additional regularization terms [28, 29], or the extension of the neural network architecture for new tasks [30, 31]. The area is advancing fast, and CL was also previously applied to solve time series problems. In [32] a new method was introduced for multi-sensor time series based on task-specific generative models and classifiers. The authors of [33] proposed a new method based on Variational Auto Encoders and Dilated Convolutional NNs aimed at multivariate time series anomaly detection problems. The work of [34] studied CL with recurrent NNs by empirically evaluating existing algorithms and introducing new benchmarks for CL with sequential data.

2.3. The Hoeffding decision tree

Standard tree-based classification or regression algorithms are not well suited for online learning [35, 36, 37]. They learn by splitting the data into separate manifolds that contain either homogenous samples from a single class or low target variable variance. For this reason, to update the structure of the tree is not as easy as in the case of NNs. Domingos and Hulten

8

introduced a new method that allows tree structure modification [38]. It uses the Hoeffding bound to calculate the number of samples in the node necessary to decide with a high probability that the best attribute provides a better split than the second-best attribute. They also showed that trees optimized using an online schema are asymptotically similar to trees learned offline using the whole data. Another tree variation introduced fast splits with non-zero information gain [39]. Forecasting capacity of Streaming Random Forest with Hoeffding trees as base models was also investigated [40]. Ensemble methods using Hoeffding trees were also investigated [41]. A parallel version of a Hoeffding tree was also developed, it computes information gain on multiple computational nodes decreasing the overall computing time [42].

2.4. Change point detection

Let $(X_t)_{t \geq 0}$ be a time series of independent and identically distributed (i.i.d.) observations. In a non-stationary environment, it may occur that the probability distribution of X changes over time, then there exists a point $t \in \{0, 1, ...\}$ such that the underlying distribution of $\{..., X_{t-2}, X_{t-1}, X_t\}$ is different from the distribution of $\{X_{t+\Delta}, X_{t+\Delta+1}, ...\}$, for a $\Delta > 0$. A change point detection task is based on minimizing a cost function that should evaluate the homogeneity of the found segments, and they are usually linked to parametric probabilistic models (e.g., using maximum likelihood method) or non-parametric approaches (as rank-based algorithms) [43]. Another critical component is a search method that could return optimal or approximate solutions [44]. Some algorithms also employ penalty terms, which are responsible for minimizing overfitting [45]. Older methods that require knowing the number of change points may not be practical in real-world applications.

Multiple approaches have been proposed to solve the problem of detecting change points [46]. However, the most well-known method seems to be the Pruned Exact Linear Time (PELT) [47] algorithm. That it can be seen as an improvement of the Optimal Partitioning method [48]. It minimizes the cost function:

$$\sum_{i=1}^{M+1}[C(y_{\tau_{i-1}+1}) + \beta],\tag{1}$$

where τ is a change point and M is the total number of change points, β is a penalization parameter.

The partitioning procedure uses a pruning method that improves computational efficiency by reducing the number of change points managed in a single iteration. There are a plethora of other change point detectors that have been proposed so far such as At Most Single Changepoint (AMOC), Binary Segmentation algorithm [49], or Bayesian Online Changepoint Detection [50].

Interest in these methods is not waning, and attempts are being made to develop change point detection methods for increasingly complex problems, including taking into account not only the non-stationarity of time series but also the fact that nowadays, more and more tasks could be multidimensional [46]. An interesting proposal is the WATCH algorithm that allows change point detection for multidimensional time series [51]. The solution is based on the analysis of distances between distributions using the Wasserstein metric. Also, interesting proposals using specific penalty functions may be found. Several works provide a recipe for estimating an interval with a useful penalty value – see, for example, the CROPS algorithm [45] or the ALPIN scheme for optimal penalty setting [44].

3. Proposed methodology

Existing research in the energy consumption forecasting in general is predominantly focused on static or non-adaptive models, overlooking the importance of continual adaptation [52, 53, 54]. Our methodology addresses this gap by emphasizing the challenges and methodologies associated with the continual improvements of forecasting models in the context of energy consumption. Moreover, we put an emphasis on the integration of change point detection methods. Recognizing that energy consumption is characterized by sudden changes throughout the year, we highlight the critical role of change point detection in enhancing the adaptability of forecasting models. By incorporating change point detection methods, our approach goes beyond traditional static models, allowing identification of significant shifts in consumption patterns.

A general methodology for multistep forecasting of multivariate time series with CL capabilities employing time series in the form of data stream processing is depicted in Fig. 1. The methodology consists of a modular schema which forms the following pipeline:

Change point detection procedure. This step is optional and is used in the Multi Collection scenario. This step aims to divide the time series into several segments. Each segment should have different statistical properties (e.g., trend direction, variance, or mean). Thus, each segment will use the different model collection for forecasting. The purpose of diving the time series into segments is to simplify the forecasting task, as each model collection should be trained on data with similar properties. Change points can be set manually by knowledge regarding the domain (e.g., based on knowl-

edge about the heating season in the natural gas consumption case) or by a change point detection algorithm such as PELT or Binary segmentation. The detected change point locations in a chosen period (e.g., the first year of the data) are then re-used in subsequent periods.

Model collection setup. The defined approaches may utilize one or more forecasting models. The forecasting model, in general, is a mathematical representation designed to make predictions about future values in a short, medium, or long-term time horizon based on historical data and patterns, for example, to forecast n hours ahead of gas consumption or m days ahead of electricity load. The first one is a Single Model Collection approach, which does not divide the data into segments but uses only one model collection for the whole dataset. The model collection is an ordered set of partial models which provides forecasts for the defined forecast horizon (for the number of steps after the forecast origin). The number of models (m) in the collection is equal to the length of the forecast horizon. The reason for using direct forecasting with multiple models instead of the cumulative approach, which reuses the forecasts from the previous timestep in predicting the next value, is to prevent the accumulation of errors through the forecast horizon. The second approach uses Multi Model Collections. The number of collections (k) depends on the number of segments into which the time series was divided, thus each segment has its own assigned model collection, i.e. if we divide the time series into k segments, k model collections will be used, one for each segment. The rationale behind this approach is that there should be a benefit in having different strategies for distinct parts of the timeseries as each segment might have different properties. The number of partial forecasting

models is the same as in the Single Model Collection approach. The specific model used is a decision made during the application of the methodology. The model has to support incremental learning, so models such as Hoeffding Tree or Stochastic Gradient Tree can be used for this task.

Model collection selection procedure. After the model collections are set, the only step left before the training phase is a logic for model collection selection. The simplest case is a Single Model Collection approach as it uses only one collection, and thus this step could be skipped. For Multi-Model Collections, there are two approaches for the collection selection, also called model aggregation scheme. In simple terms, the model aggregation scheme is the plan for deciding which set of prediction models to use based on the changes observed in the time series data. The first one, called the Pure Change Point-Divided schema, selects the model collection based on the current time series segment (segments are divided by change points from the *Change point detection* step). When a change point occurs, the current model collection is exchanged for the subsequent one. The rationale behind this schema is that when the time-series distribution shifts, then it is beneficial to switch to a different group of forecasting models. The second approach, called Mixed Change Point-Divided schema, is based on the hypothesis that the change points do not occur at the exact same location in every period, but the location varies in some close subsegments. Thus, this approach works with a boundary of b steps, which defines a segment of length $2b$ (b step prior to the change point and b steps after it). Both model collections (before and after the change point) are used inside this boundary. They are trained on the new instances and produce forecasts. A forecast error measure E is

computed using an arbitrarily selected metric (e.g. MAE, MSE, etc.) for both model collections. Error calculation is used at timestep $t + 1$ for the final forecast construction, and this process can be performed in one of two ways:

1. **WAVG procedure:** Let R_{WA} be a collection of m forecasting models generated by the WAVG procedure $\left(R_{WA}^{(t)} = (r_1^{(t)}, r_2^{(t)}, ..., r_m^{(t)})\right)$. Furthermore, we denote by C_1 and C_2 two disjoint collections of partial models $R_{C_1}^{(t)}$ and $R_{C_2}^{(t)}$ before and after a specific change point. The forecast value of model r_i at timestep t $(f_i^{(t)})$ is calculated as a weighted average of the forecasts produced by the following two model collections $R_{C_1}^{(t)}$ and $R_{C_2}^{(t)}$:

$$f_i^{(t)} = \frac{w_{C_1} f_{i,C_1}^{(t)} + w_{C_2} f_{i,C_2}^{(t)}}{w_{C_1} + w_{C_2}}, \tag{2}$$

where the weights w_{C_1} and w_{C_2} are defined according to the previous performances (errors $E_{C_1}^{(t-1)}$ and $E_{C_2}^{(t-1)}$):

$$w_{C_1} = 1 - \frac{E_{C_1}^{(t-1)}}{E_{C_1}^{(t-1)} + E_{C_2}^{(t-1)}},$$

$$w_{C_2} = 1 - \frac{E_{C_2}^{(t-1)}}{E_{C_1}^{(t-1)} + E_{C_2}^{(t-1)}}.$$

2. **SWITCH procedure:** Let R_{SW} be a collection of m forecasting models generated by the SWITCH procedure $\left(R_{SW}^{(t)} = (r_1^{(t)}, r_2^{(t)}, ..., r_m^{(t)})\right)$. The forecast at the timestep t is also based on the forecasts produced by the two model collections, as in the WAVG procedure case. However, in this procedure, only the forecast at the timestep t by the model collection with lower E at the timestep $t - 1$ is used, thus the models switch during the boundary period. The forecast value of model r_i

14

$(f_i^{(t)})$ at the timestep t is produced by the partial model of the selected collection C_1 or C_2:

$$f_i^{(t)} = \begin{cases} f_{i,C_1}^{(t)}, & \text{if } E_{C_1}^{(t-1)} < E_{C_2}^{(t-1)}, \\ f_{i,C_2}^{(t)}, & \text{otherwise.} \end{cases} \tag{3}$$

In both cases, these procedures help to decide how to select different prediction models based on their past performance around detected change points in the data. WAVG combines predictions from different models with weighted averages, giving more influence to historically accurate models. SWITCH, on the other hand, selects predictions from one model based on its past performance, adapting to changing circumstances in the data. After the boundary sub-segment ends, only a single model collection is used, and its choice is the same as in the Pure Change Point-Divided schema case.

Model training. When the model collections are set up and the selection schema is selected, the model collections could be used in a training phase. The methodology is meant for data stream incremental learning; thus, the models are trained after each data instance. Before training, the models produce the forecasted values and error metrics are computed and used if the Mixed Change Point-Divided schema is used.

4. Experiments

The conducted experiments aim to verify the accuracy and efficiency of the proposed method on real-world data and evaluate the defined model collection selection schemas as beneficial over naive baseline models. We intend to answer the following research questions:
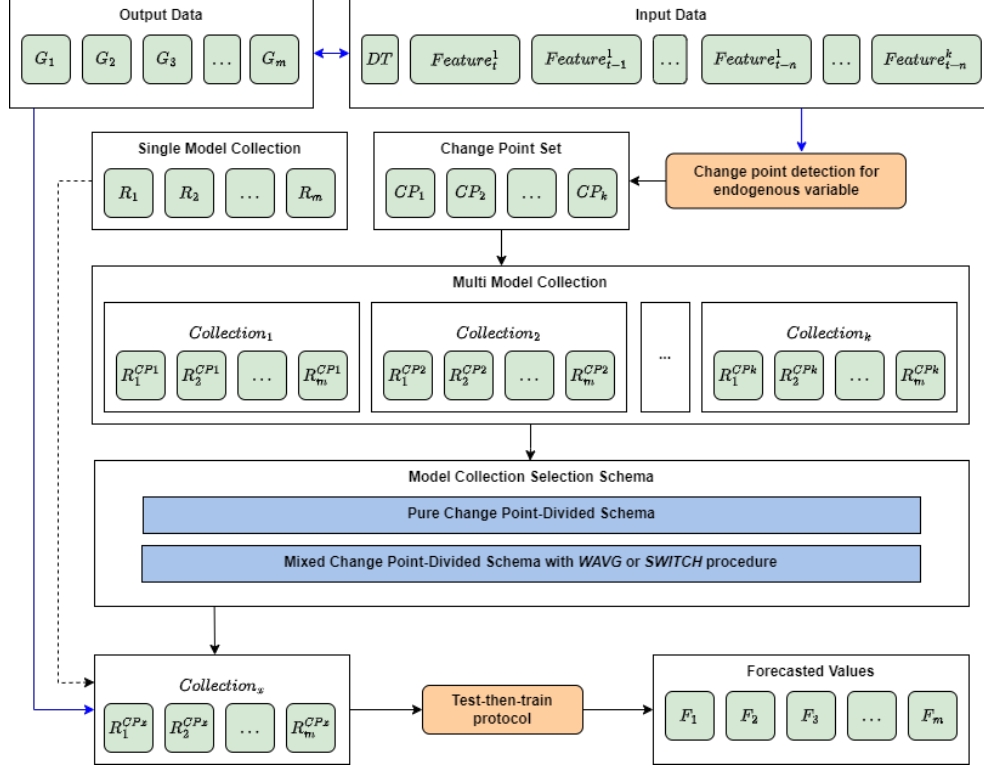
Figure 1: The proposed multistep forecasting pipeline with CL support.

- **RQ1:** How does the performance of a single-model continuous learning approach (SMCA) compare to a multimodel approach aggregated by year quarter (QDMDC) in the context of real-world data?

- **RQ2:** May we benefit from the model aggregation scheme based on change points detected by the PELT algorithm compared to the quarter aggregation?

- **RQ3:** Does the number of detected change points affect the accuracy of the forecast?

- **RQ4:** Is there an advantage in employing forecast ensembling (MCPDMC-

WA) or model switching (MCPDMC-SW) in proximity to change points, with the main consideration being the Symmetric Mean Absolute Percentage Error (SMAPE) metric?

4.1. Experiment setup

Dataset. The data covers eight years, from January 1, 2013, to December 31, 2020, with data available at an hourly frequency [55][1]. It consists of 70,104 data points, compiled from three main components. The first component comprises consumption data, focusing on Prague, the capital of the Czech Republic. The Prague distribution network served 422,926 customers in 2018, with a total consumption of 3.82 billion m3. The second component incorporates weather variables obtained from the Prague LKPR airport weather station. These variables are derived from periodic METAR reports issued by airports and are archived for long-term preservation. The third component that represents economic characteristics is natural gas price data. We have obtained price data from the Czech Energy Regulation Office and included them in the dataset. The natural gas consumption time series is shown in Fig. 2, as well as the temperature (Fig. 4), which is the most important exogenous variable, since gas consumption is highly dependent on the outside temperature because of the household heating.

Natural gas consumption over multiple years exhibits a clear cyclical phenomenon, as shown in Fig. 2. Cycles reflect seasonal variation of the outside temperature, affecting heating and cooling demand. Fig. 4 illustrates the relationship between consumption and temperature, indicating that consump-

---

[1]The dataset is available online: `https://ai.vsb.cz/natural-gas-forecasting`

Table 1: Description of the features included in the natural gas dataset.

| Feature name	Description	Measuring unit
Year	2013-2018	-
Month	1-12	-
Day	1-31	-
Hour	0-23	-
Day of week	1-7	-
Before holiday	Is the next day a holiday?	-
Holiday	Is the current day a holiday?	-
Temperature	Air temperature at a 2-meter height above the earth's surface	°C
Pressure	Atmospheric pressure at weather station level	mm Hg
Pressure (sea level)	Atmospheric pressure reduced to mean sea level	mm Hg
Humidity	Relative humidity at height of 2 meters above the earth's surface	%
Wind direction	Mean wind direction at a height of $10-12$ meters above the earth's surface over a 10-min period	compass points
Wind speed	Mean wind speed at a height of $10-12$ meters above the earth's surface over a 10-min period	m/s
Phenomena	Special present weather phenomena observed at or near the aerodrome	Text value
Recent phenomena	Recent weather phenomena of operational significance	Text value
Visibility	Horizontal visibility	km
Dewpoint	Dewpoint temperature at a height of 2 meters above the earth's surface	°C
Cloud cover	Total cloud cover	Text value
Price	Weighed average	EUR / MWh
Consumption	Consumption of the distribution network for the current hour	m^3/h
Temperature YRNO	Forecasted temperature	°C
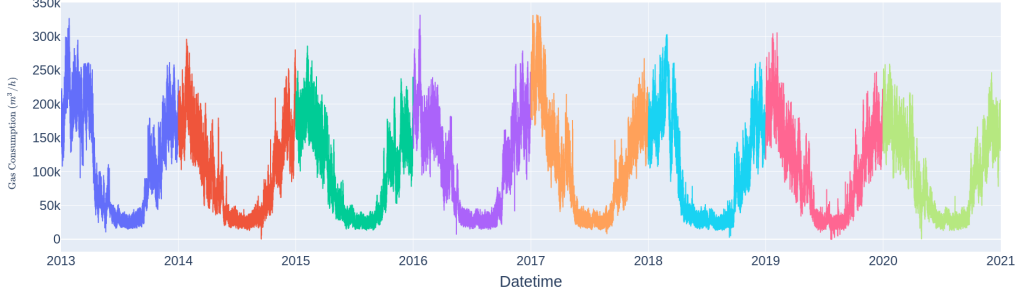
Figure 2: Natural gas consumption (endogenous variable) from January 1, 2013, to December 31, 2020 (years are divided by colors).
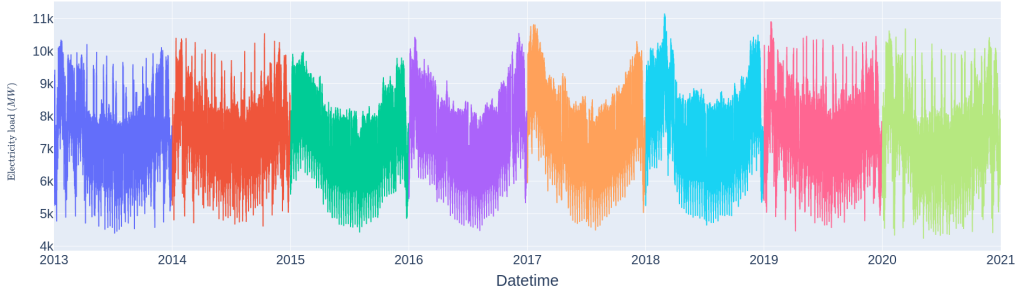

Figure 3: Electricity load (endogenous variable) from January 1, 2013, to December 31, 2020 (years are divided by colors).

tion is high during the cold and low during the warm months. Consumption also shows short-term trends corresponding to transition periods between the heating and summer seasons, when the temperature fluctuates rapidly.

To investigate the ability to generalize of the proposed methodology, we have conducted the experiments also for the electricity load data. The electricity load forecasting experiments utilized a dataset comprising raw load data spanning the same time frame as the natural gas dataset, covering the period from January 1, 2013, to December 31, 2020, with hourly granular-
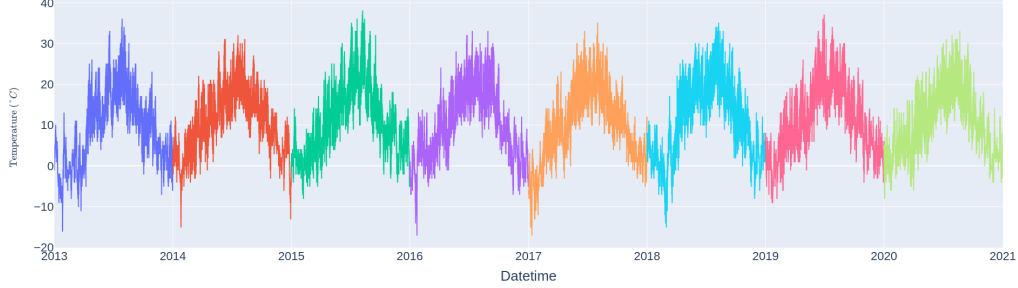
19

Figure 4: Temperature (exogenous variable with the strongest correlation coefficient (-0.871) to the natural gas consumption) from January 1, 2013, to December 31, 2020 (years are divided by colors).

ity. Weather data and calendar information were sourced from the natural gas dataset, as they were measured at the same location, namely the city of Prague. However, due to insufficient quality, the price component was excluded from this dataset. The electricity load time series is depicted in Fig. 3

Fig. 5 shows the change points detected by the PELT algorithm (*Low* settings) for three selected years of the natural gas consumption data. The detected change points partition the data into segments with different levels or trends of gas consumption. Segments overlap partially, with better agreement during the summer and fall months, when consumption increases. Spring months have more frequent levels and trend changes, resulting in lower segment overlap. The same procedure were applied to the electricity load data (see Fig. 6). We can see that the properties are similar to the natural gas data, although the ; however, we can see that the daily/weekly electricity load patterns evolve over time, the volatility of the time series is lower, and the pattern became more stable, the difference can be clearly

20
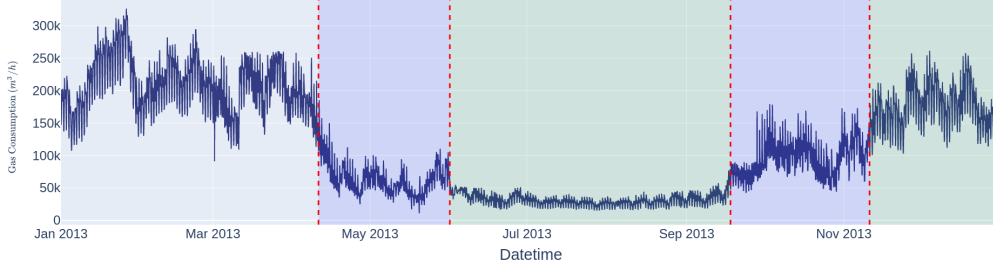
(a) Change points detected in 2013 data.
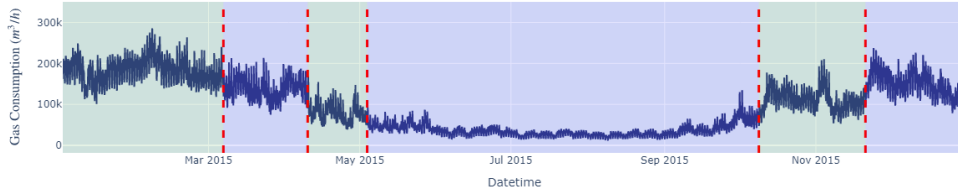
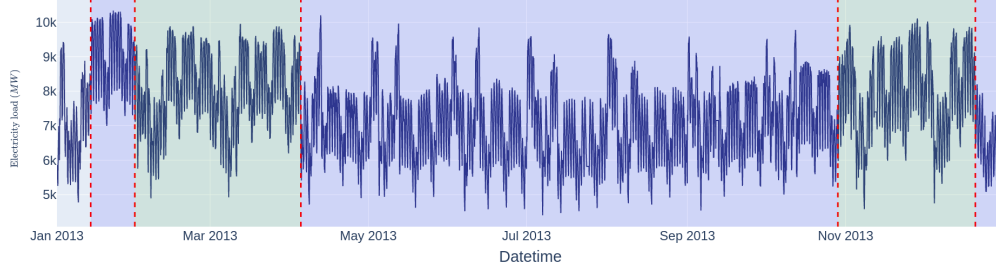(b) Change points detected in 2015 data.
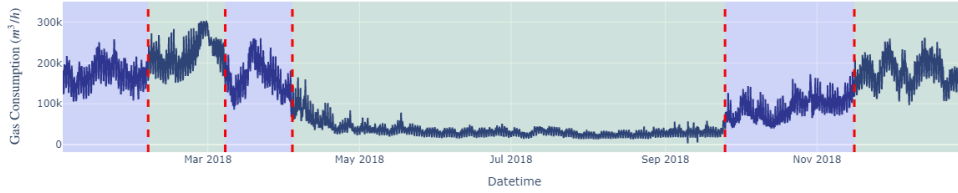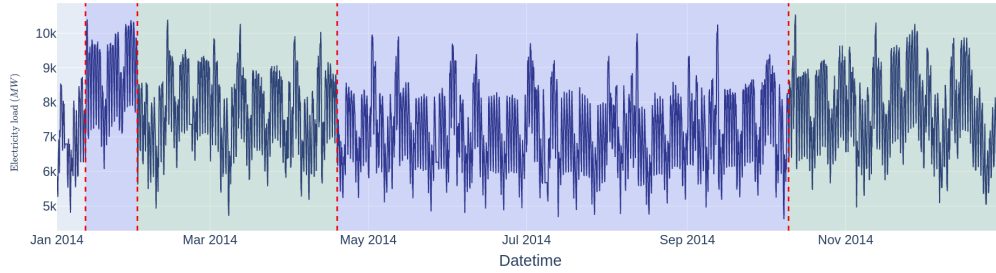

(c) Change points detected in 2018 data.

Figure 5: Detected change points by PELT algorithm with *Low* settings in selected years of the natural gas consumption data.

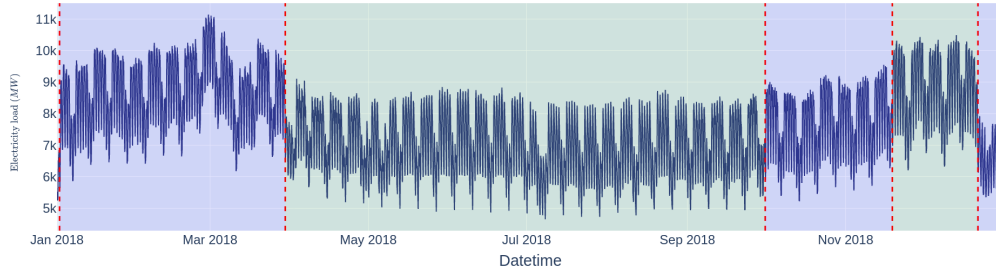spotted if we compare the patterns depicted in Figures 6a and 6c.

Dataset preprocessing. The original data is transformed into a suitable form for machine learning algorithms and divided into sets of pairs of input

(a) Change points detected in 2013 data.

(b) Change points detected in 2014 data.

(c) Change points detected in 2018 data.

Figure 6: Detected change points by PELT algorithm with *Low* settings in selected years of the electricity load data.

and output vectors. An input-output tuple is created for each forecast origin (i.e., a point in the time series from which the forecast is done, for example,

midnight). The input vector includes past data up to the forecast origin, and the output vector includes natural gas consumption values for each hour of the following day. The input vector consists of the so-called lagged features - values obtained prior to the forecast origin. In our experiments, these are Consumption (endogenous variable) and Temperature, Price, Wind Speed, and Humidity (exogenous variables) of the original dataset. We used a fixed window of 72 hours for these lagged variables. On top of that, the input vector contains the Temperature Forecast [2] for the next 24 hours and usual date time variables such as Day of week, Month or if the day is a holiday. The forecast origin is at midnight, the forecast horizon is 24 hours long; thus, the output vector consists of the Consumption variable values for the next 24 hours, and each forecasted value corresponds to one hour of the next day.

Algorithms used in the forecasting pipeline. The methodology proposed in Section 3 is generally defined without the need for specific change point detection and forecasting algorithms. For the experiments, we have used a PELT algorithm for the change point detection phase, as it is probably the most known state-of-the-art algorithm in this domain (see 2 for more details) and Hoeffding Tree as a forecasting model due to the fact that it supports incremental learning out of the box and tree-based algorithms performed very well on the dataset used in classical offline learning scenarios [55]. The Hoeffding Tree Regressor with 7 instances (one week of the input data) a leaf should observe between split attempts (the so-called grace period), decay of the model selector of 0.2 and threshold (τ) below which

---

[2]Forecasted by YR.no (`https://www.yr.no`)

a split is forced to break ties of 0.5 is used as a forecasting model. The parameter values are based on the hyper-parameter tuning procedure employing the first year of the data. As the defined multistep forecasting task requires multiple output values, the multimodel direct forecasting approach is used. There is a sequence of 24 models; each value of the output vector is forecasted by its corresponding model. To be able to compare the Hoeffding Tree Regressors models also the experiemnts employing deep learning models were carried out. Three variants of the deep learning model were used, each model in three configurations. The general architecture of the model is depicted in Fig. 7. The model variant depends on the layer type used as its input; models with fully connected, LSTM or GRU layer were used. Three sets of layer sizes were used, large, medium, and small, the neuron counts in the layers are also stated in Fig. 7. PELT algorithm with the $L2$ segment model, 168 hours of minimum segment length, and a subsample rate of 24 hours is used for the change point detection phase of the process.

The computational complexity of the suggested methodology depends mainly on the Hoeffding tree regressor implementation. In general, the Hoeffding Tree regressor is generally considered to be linear with respect to the number of training instances (n) and logarithmic with respect to the number of features (d) [56]. During the training phase, the algorithm incrementally builds a decision tree as new instances arrive. At each step, it evaluates the splitting criteria for each candidate attribute to determine the best split. This process involves computing information gain or other statistical measures to assess the quality of the splits. The preprocessing steps use the linear transformation of the attributes and do not add complexity. The

24

most computation complexity is then in forecasting using a deep learning approach, especially LSTM-based models. The computational complexity of Long Short-Term Memory (LSTM) networks depends on several factors, including the number of input features (n), the number of LSTM units or cells (m), the length of the input sequence (t), and the number of output units (k). The computational complexity of a single LSTM cell operation (including all gates and computations of cell state) is typically $O(m^2 + m * n)$, where m is the number of LSTM units and n is the number of input features. If the input sequence has length t and there are k output units, the total complexity of a single forward pass is approximately $O(tm^2 + tmn + tm * k)$ [57]. For deeper LSTM networks with multiple layers, the computational complexity increases with the number of layers. If there are L layers in the network, the overall complexity becomes $O(Ltm^2 + Ltmn + Ltmk)$. The computational complexity of the PELT (Pruned Exact Linear Time) algorithm depends on several factors, including the length of the time series data (n), the number of change points detected (m), and the desired level of accuracy. The PELT algorithm is known for its computational efficiency, especially compared to other change-point detection algorithms. It operates in linear time complexity, meaning that its computational cost increases linearly with the size of the input data [47]. Our proposed methods add only a tiny part of the computational complexity, resulting in a solid improvement in forecasting precision.

Experimental protocol. Because we employ data stream processing, thus only the current data instance is available to the model at each training step. Models are trained incrementally daily. The interleaved-test-then-train

25

Figure 7: Deep model architecture used in the experiments. The scheme shows the general architecture, the first layer can be either Dense, LSTM or GRU layer, it depends on the selected variant of the model. Three sizes of the deep learning model were tested, Large (L), Medium (M) and Small (S), size of the network affects the number of neurons in the first two layers. Number of neurons in each of the variant is stated in the diagram.

evaluation [58] was used in all experiments. It works with a data stream of instances, each instance is first used for the inference (testing phase), then it is used for the model training. allowing for the gradual refinement of model accuracy. This strategy allows a gradual refinement model's accuracy. Following this sequential evaluation process, the model consistently faces unseen samples. This approach offers the benefit of eliminating the need for a separate holdout set for testing, ensuring optimal utilization of the available data [59].

The following experiments were carried out:

- **Single Model Collection Approach (Baseline).** Only one model collection is used for the whole data stream. The model collection consists of 24 forecasting models, each model being used for one hour of the day.

- **Quarter-Divided Model Collection.** Each quarter of the year has its model collection. The collection choice depends on the current quarter of the year.

- **Pure Change Point-Divided Model Collection.** The year is divided into multiple segments by the change points in gas consumption detected by the PELT algorithm. Each of the segments has its own collection of models. The change points are detected using only the first year of the data (2013), and the same change point locations are reused in the next years. When the change point occurs, the model collections are switched.

- **Mixed Change Point-Divided Model Collection.** The approach is the same as the pure one at its core; however, there is a 7-day period before and after each change point. Two strategies for forecast ensembling in these periods are tested. During these 14-day long periods, both model collections are trained on the data. The first strategy uses weighted forecast averaging. The Mean Absolute Error (MAE) is computed for both models, and it is used to calculate the weighted average of the two models forecast in the next forecast step. The model with a higher error in the previous step has a lower weight in the average and vice versa. The second strategy uses model switching. The MAE is computed for both models as well, but only the forecast produced by the model with the lower error in the previous step is used. Outside these periods, the process works in the same way as in the Pure Change Point-Divided Model Collection case.

The Pure and Mixed Change Point-Divided Model Collection experiments were carried out with three different PELT algorithm settings, which detected 4 (*Low* settings), 7 (*Medium* settings), and 13 (*High* settings) change points; to be able to assess the effect of increasing the number of detected change

points (i.e. number of model collections) on the forecast error. The limitation of the PELT algorithm settings to only three options in our study was a deliberate choice based on practical considerations. While it's possible to introduce a more varied range of settings for the PELT algorithm, we opted for a more focused approach to streamline our analysis. In reality, the differences among various settings of the PELT algorithm might be relatively small, especially concerning their impact on performance. Therefore, instead of exhaustively exploring a wide range of settings, we adopted a scenario-based approach. By utilizing three distinct settings, we aimed to assess their effects on performance in a simpler manner. This simplified approach allowed us to effectively evaluate the performance of the algorithm under different configurations without overly complicating the analysis. While it's possible to introduce more varied settings for the PELT algorithm, our approach provided insights into its behavior and effectiveness using the three defined scenarios. The experimental setting that we designed for validating the proposed methodology is depicted in Figure 8.

Result evaluation. Let G_t be the ground truth of natural gas consumption or electricity load and the variable F_t the forecast value, both at time step t. We have used the standard set of commonly used error metrics, each summarizing different information, [60], *MAE, MSE* and *SMAPE* defined in expression (4), (5) and (6), respectively togerther with a Diebold-Mariano test [61].

$$MAE = \frac{\sum_{t=1}^{n} |G_t - F_t|}{n}, \tag{4}$$
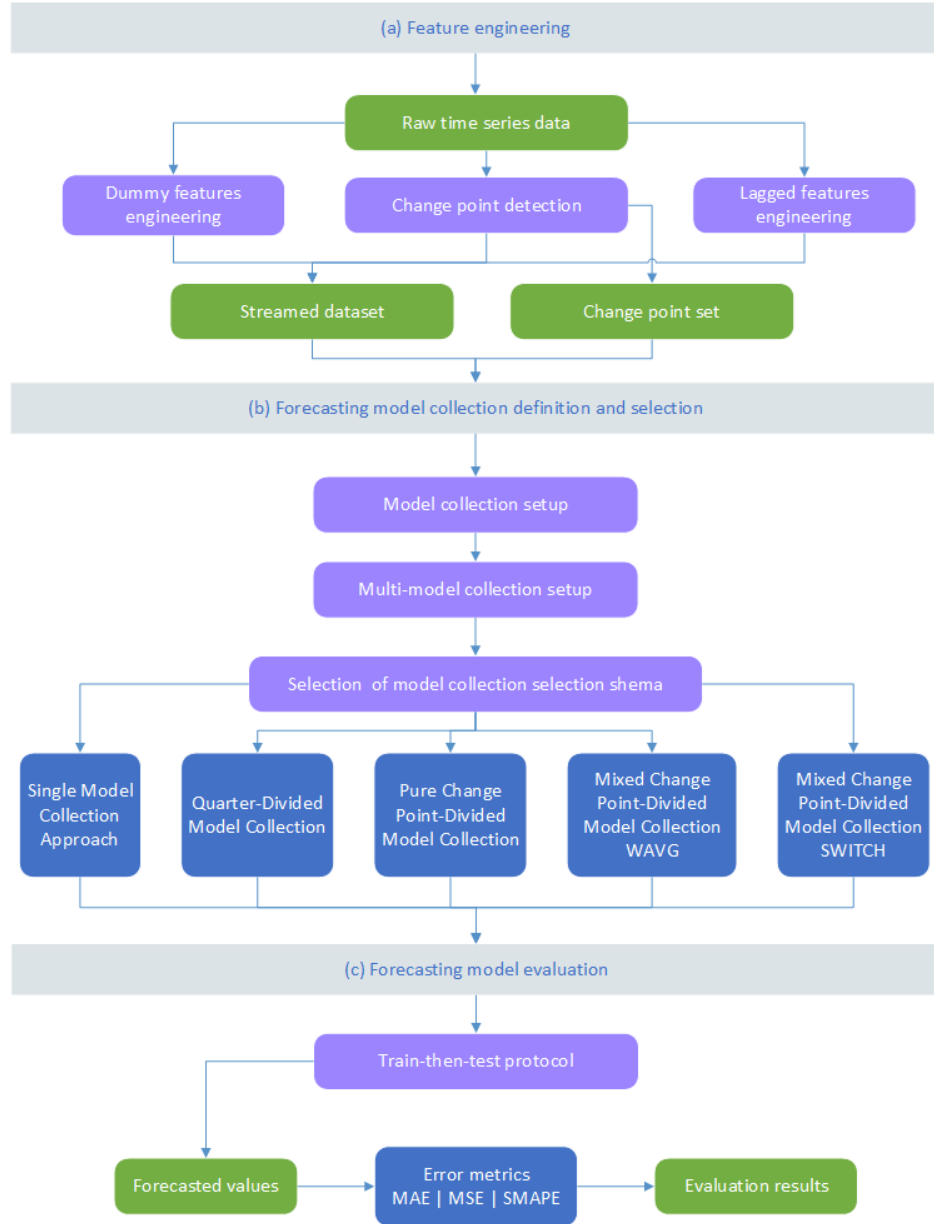
$$MSE = \frac{\sum_{t=1}^{n} (G_t - F_t)^2}{n}, \tag{5}$$

28

Figure 8: Designed experimental pipeline overview with high level of abstraction.

$$SMAPE = \frac{100}{n} \sum_{t=1}^{n} \frac{\mid G_t - F_t \mid}{\frac{1}{2}\left(\mid G_t \mid + \mid F_t \mid\right)}. \tag{6}$$

The results are evaluated for the years 2014 to 2020. The first year 2013 was intentionally left out of the evaluation process because the data were used in the change point detection phase for the Pure and Mixed Change Point-Divided Model Collection experiments; thus, it could be taken as an information leak and this phenomenon would distort the comparison with other approaches in this year.

Implementation and reproducibility. The experiments were carried out in Python (v. 3.11) using the Ruptures library (v. 1.1.7) for the PELT algorithm and the River library (v. 0.15.0) for the Hoeffding Tree Regressor. Tensorflow2 (v. 2.15.0) framework was used for the training of the deep learning models. Pandas (v. 1.5.3) and Numpy (v. 1.24.2) libraries were utilized for dataset manipulation and preprocessing. The codebase with the method and experiment implementation is available in the git repository[3]

5. Results

Two baseline and three models based on the proposed methodology were experimentally evaluated together with three deep learning model architectures in three different settings to be able to compare the Hoeffding Tree-based models to other approaches. Both baseline models omit the change point detection integration, as the first uses only one model collection for the data stream, and the second aggregates the model collections using year quarters. Models based on the proposed methodology were evaluated using

---

[3]`https://github.com/rasvob/Hoeffding-Trees-with-CPD-Multistep-Forecasing`

three levels (low, medium, and high) of change point density. For each model, a Mean Absolute Error (MAE), Mean Square Error (MSE), and Symmetric Mean Absolute Percentage Error (SMAPE) were calculated. The aggregated results are summarized in Tables 4, 5, 6, 7 for the natural gas consumption data and in Tables 8, 9, 10, 11 for the electricity load data. Figures 12 and 13 provide a detailed view of the results that provides insight into the model accuracy differences is included in Figures 14, 15 and Tables 2, 3.

5.1. Threats to validity

In the realm of empirical research in the domain of natural gas consumption, respectively electricity load, forecasting, the identification and mitigation of validity threats play a crucial role in ensuring the reliability and applicability of predictive models. Our goal is to create accurate forecasts capable of informing real-world decision making, and it becomes imperative to recognize potential challenges that could compromise the validity of our results. The thoughtful consideration of potential threats becomes the cornerstone of building robust, accurate, and actionable forecasting systems. We explore specific threats to validity, both internal and external, in the context of our forecasting task and outline strategies to mitigate these challenges. We consider the following threats to be the most severe in terms of our research.

5.1.1. Data selection bias

Connection to the research. Data selection bias occurs when data collection is collected from a biased and non-representative population. In our context, data selection bias could manifest itself if the data set predominantly represents certain seasons, leading to skewed patterns in natural gas

consumption. For example, if the data set is predominantly from colder seasons, the model may not generalize well to warmer seasons, affecting the reliability of the forecasting system.

Mitigation strategy. To address data selection bias, we ensured that the seasons are equally represented; the same applies to any medium- or long-term trend, using a dataset that contains data from 2013 to 2020. This approach aims to enhance the model's adaptability to various consumption patterns and ensuring better generalization across different scenarios. Beside the natural gas consumption data, the experiments were also carried out using an electricity load data to ensure that the proposed methodology can be adopted in different tasks and is not focused solely on a single dataset.

5.1.2. Methodological bias

Connection to the research. Methodological bias refers to bias introduced by researchers in the methodological design of the study. Both machine learning and change point detection algorithms may be sensitive to hyperparameter settings. The selection of the number of used change points and the way to deal with them may introduce potential sources of methodological bias because it may severely affect the results obtained. Moreover, from the point of view of the experiment design, there is a risk of selection bias in the evaluation metrics and selection of the train-test set as well.

Mitigation strategy. Sensitivity analyzes were performed to evaluate the impact of various methodological choices on the performance of the model. Several different configurations for the number of change points used were tested and multiple approaches to deal with them were used during the experiments. The results were analyzed in detail and the different settings and

approaches were compared with each other and also with respect to different time periods of the year. The proposed Hoeffding Trees-based models were moreover compared to a wide variety of deep learning models. Cherrypicking evaluation metrics that highlight the model's strengths while downplaying its weaknesses was avoided by reporting multiple widely used error metrics, such as MAE, MSE, or SMAPE. The interleaved test-then-train protocol was used during the training and evaluation of the models to assess generalization performance. This ensures that the model performance is not overly tailored to the peculiarities of the selected parts of the data. This approach was designed to minimize methodological bias and ensure that the methodologies chosen produced robust and reliable results.

5.1.3. Confounding features threat

Connection to the research. Confounding features threat arise when external factors (exogenous variables) that could affect predictions are not considered in the methodological design. External factors, such as weather, especially temperature and forecast for the next day, or date-time features can influence natural gas consumption. Failure to account for these confounding features and rely solely on the endogenous variable during the methodological design could lead to inaccurate predictions.

Mitigation strategy. To address the threat of confounding features, we have conducted extensive research in this domain in our work [55] to identify potential external factors that impact gas consumption. Relevant exogenous variables were incorporated into the model.

5.2. Methodology and results in context

As it was described in Section 2, many existing studies are based on conventional datasets [10, 62]. In contrast, the proposed research uses a unique and underexplored real-world dataset [55] in a CL context, introducing a new perspective to the field of natural gas forecasting. Although the scarcity of prior research on this specific dataset contributes to the originality of our work, it also poses distinct challenges in the comparison of related work results from the forecast accuracy point of view. Due to this fact, we have decided to put our methodology and experiment results in contrast with other related methods from the perspective of model adaptability, feature engineering, approach to analysis of results, and model accuracy.

5.2.1. Forecasting model adaptability

Research context. The prevailing trend in related research within the realm of natural gas consumption, or energies in general, forecasting predominantly focuses on static models with limited adaptation capabilities [10]. Moreover, these models are typically used for single-point forecasts; thus, they provide forecasts as scalars for a single point in time [63].

Proposed methods contrasts. Unlike related approaches that focus on static models and offer forecasts for a single point in time, our research takes a different approach. We use CL principles by treating natural gas consumption, or electricity load, as an ongoing data stream. This means that the proposed models learn and adapt to new data in real time. Note that the proposed method is focused on vector forecasts with a forecast horizon beyond just one time unit ahead, thus it is capable of predicting the whole forecast horizon at once; we have focused on a 24 hours long forecast horizon

34

in our experiments as this is the common forecast horizon length in this domain.

5.2.2. Feature selection possibilities

Research context. Related research primarily centers on leveraging statistical methods and neural networks for natural gas consumption forecasting. However, we observe that statistical models often rely on univariate time series, such as Autoregressive integrated moving average (ARIMA) [64], which limits their ability to capture intricate relationships within multivariate data. Similarly, neural networks, while powerful, present limitations for modeling time series in contexts when the model should adapt according to changes of the input data distribution. Typically, longer input sequences are required, offer limited opportunities for feature engineering through domain knowledge, and generally focus on offline learning approaches [65]; therefore, models often rely on the availability of the entire dataset during training. In particular, conventional practice in the energy domain involves implementing a single forecasting model for the entire dataset.

Proposed methods contrasts. In contrast to the prevalent emphasis on statistical methods and neural networks in natural gas consumption, or electricity load, forecasting, our research prioritize tree-based models for vector forecasts. Although these models are conventionally designed for scalar output, the methodology proposes a model collection setup which enables the models to be used for tasks which require vector output. Unlike statistical models and neural networks, tree-based models present robust opportunities for feature engineering, allowing us to make use of multivariate data more effectively. A previous study showed that tree-based models achieve bet-

ter accuracy compared to statistical methods [55], however, it is important to note that the research was focused on feature engineering in an offline learning context. The approach presented in this article capitalizes on the recurrent behavior presented in the consumption of natural gas. However, to ensure that tree-based models provides the mentioned advantages compared to other methods, a series of deep learning experiments were conducted as well, thus it is possible to compare the different approaches.

5.2.3. Results analysis

Research context. The prevailing trend in related research predominantly centers around the analysis of overall forecasting accuracy, where models are typically trained only once. Although this approach provides valuable insights into the general performance of forecasting models, it often overlooks time-related patterns within the time series data. Additionally, results analysis in the field tends to focus on classical properties, examining only statistical aspects of error distribution throughout the day of the week or month, and these analyses often operate under the assumption of a static forecasting model.

Proposed methods contrasts. While we cover the standard analyses like overall forecasting accuracy and classical error distribution, our research stands out by giving a strong focus to how our models learn online. We examine how the accuracy of the forecast changes over time, examining how the number of detected change points impacts the predictions. We are especially interested in how our forecasts differ in detail during periods near these change points. This approach allows us to understand how well our models adapt to the evolving nature of natural gas consumption patterns.

5.2.4. Forecast accuracy

Research context. Given the rather novel dataset used and the prevailing focus of related energy forecasting research on offline learning and single-value-ahead forecasts, direct numerical comparisons with existing results could be misleading. We acknowledge the difficulty in numerical comparisons and have implemented specific measures to address this concern, ensuring the robustness of our results. It is important to note that from our previous research [55], we can assert that the Symmetric Mean Absolute Percentage Error (SMAPE) for 24-hour ahead forecasts for this dataset generally falls within the range of 7% to 15% for models trained offline. This variation depends on the method employed and its complexity.

Proposed methods contrasts. To navigate the complexities arising from the novelty of our dataset and the lack of continual/online learning research within this domain, we opted to establish two baseline models, SMCA and QDMDC. By defining these baselines ourselves, we ensured a reliable benchmark for comparison against more intricate approaches. In particular, there has been no prior research specifically focusing on continual or online learning for this dataset. Using proposed approaches that incorporate change point detection, we successfully reduced the Symmetric Mean Absolute Percentage Error (SMAPE) by approximately 5 to 6.5 % compared to the baseline models. Importantly, this improvement was achieved with minimal additional computational overhead, which is particularly evident when compared to the QDMDC approach. This nuanced comparative analysis underscores the effectiveness of our methodology in enhancing forecast accuracy while maintaining computational efficiency.

6. Lessons learned

We have grouped conclusions according to the research questions formulated in Section 4 to facilitate a systematic evaluation of the results obtained.

RQ1: How does the performance of a single-model continuous learning approach (SMCA) compare to a multimodel approach aggregated by year quarter (QDMDC) in the context of real-world data?

For the natural gas consumption data, the HT-SMCA approach exhibited a marginally lower SMAPE than the HT-QDMDC approach, a baseline that segments the data by quarters of year (Table 4). Analysis of Fig. 12 shows that this discrepancy in SMAPE was mainly due to the lower SMAPE of HT-SMCA in 2014, after the first two years of training. With more training data, the models performed similarly with negligible differences. Furthermore, the HT-QDMDC approach yielded a higher SMAPE but slightly lower MSE compared to the HT-SMCA approach (Table 4). However, the performance difference was not significant. Consequently, the straightforward HT-SMCA approach proved preferable to the HT-QDMDC approach, which naively segmented the data by quarters of years without capitalizing on inherent cyclical patterns in natural gas consumption.

For the electricity load data, the HT-SMCA has also lower error values compared to the HT-QDMDC approach (see Table 8). Fig. 13 shows the progress of SAPE over the years and we can see that the HT-SMCA approach was superior to HT-QDMDC at all times.

Deep learning based models performed worse compared to HT-SMCA

38

or HT-QDMDC in both datasets regardless the SMCA or QDMDC approach used, moreover the models employing LSTM or GRU had even higher forecasting errors compared to the fully connected model. Using just a single model for the whole dataset (SMCA) approach led to lower error compared to the QDMDC approach thus we can see that deep learning models do not benefit from the model switching as in the Hoeffding Tree-based models case.

RQ2: May we benefit from the model aggregation scheme based on change points detected by the PELT algorithm (PCPDMC) compared to the quarter aggregation?

In case of natural gas consumption dataset the change point divided model aggregation scheme (HT-PCPDMC) outperformed both baselines by all three metrics. The simplest variant (Pure) achieved the best performance in almost all PELT settings, except for the *low* setting, where it was comparable to the weighted average forecast ensemble variant (HT-PCPDMC-WA). However, the error of HT-PCPDMC-WA increased with the number of change points. Tables 5, 6, 7 and Figures 12a and 12c show the data and the evolution of SMAPE over time.

In electricity load dataset experiments the HT-PCPDMC scheme also provided lower errors compared to HT-QDMDC regardless of used number or change points in this case, however *Low* and *Medium* change point numbers provided better results compared to *High* PELT setting, as we can see in the Tables 9, 10, 11, the progress of SMAPE in different

PELT settings can be seen in Fig. 13. We can see that electricity load forecasting becomes more complex over time compared to natural gas consumption as there is an error level shift present in all models used in years 2019 and 2020, furthermore, we can see that the proposed HT-PCPDMC performed the best even after the error level shift occurred and SMAPE for year 2020 was the lowest in all three PELT settings.

The PCPDMC scheme compared to QDMDC was beneficial for the deep learning models as well, as the SMAPE was lower, we can see that lower number of change points provides better results as the SMAPE increased with the number of change points. However, we can see that deep learning models performed better if a single model was used for the entire data stream.

Fig. 9 illustrates the p-values resulting from the Diebold-Mariano test conducted on the selected models for natural gas consumption. It is evident that, with the exception of the HT-PCPDMC model utilizing the *medium* PELT settings, each model employing an aggregation scheme based on change points detected by the PELT algorithm exhibits a p-value lower than 0.05, when compared to both the baseline HT-QDMDC and deep learning-based models.

Similarly, the p-values derived from the Diebold-Mariano test for the electricity load dataset are presented in Fig. 10. Here, the comparison indicates that, across the baseline deep learning models and HT-QDMDC, the p-values for the comparison of the other models selected remain below the threshold of 0.05, except for the HT-PCPDMC model utilizing the *high* PELT settings. Thus, irrespective of the dataset ana-

lyzed, the null hypothesis that there is no difference in the accuracy of the forecast between two competing forecasting models can be rejected.

	HT-SMCA	HT-QDMDC	FCN-SMCA (M)	FCN-QDMDC (M)	FCN-PCPDMC (M, Low)	HT-MCPDMC-WA (Low)	HT-PCPDMC (Low)	HT-PCPDMC (Medium)	HT-PCPDMC (High)
HT-SMCA		0.67	9e-18	3.3e-35	1.1e-34	3.7e-07	5.7e-07	0.55	0.0002
HT-QDMDC	0.67		1.3e-18	4.5e-38	3.2e-34	3.4e-05	7.2e-05	0.41	5.1e-05
FCN-SMCA (M)	9e-18	1.3e-18		5.7e-16	6.8e-06	2.5e-23	3.1e-23	7.1e-17	7.7e-07
FCN-QDMDC (M)	3.3e-35	4.5e-38	5.7e-16		7.2e-07	2.6e-39	5.1e-39	9.3e-34	3.3e-23
FCN-PCPDMC (M, Low)	1.1e-34	3.2e-34	6.8e-06	7.2e-07		1e-41	2.2e-42	2e-31	1.1e-14
HT-MCPDMC-WA (Low)	3.7e-07	3.4e-05	2.5e-23	2.6e-39	1e-41		0.62	3.8e-05	9.5e-10
HT-PCPDMC (Low)	5.7e-07	7.2e-05	3.1e-23	5.1e-39	2.2e-42	0.62		3.6e-05	1.1e-09
HT-PCPDMC (Medium)	0.55	0.41	7.1e-17	9.3e-34	2e-31	3.8e-05	3.6e-05		1e-05
HT-PCPDMC (High)	0.0002	5.1e-05	7.7e-07	3.3e-23	1.1e-14	9.5e-10	1.1e-09	1e-05	

Figure 9: Matrix displaying $p-values$ of the Diebold-Mariano Test used for a pair-wise comparison of selection of the best performing models with the baseline approaches using both Hoeffding Trees and Deep learning models for the natural gas consumption dataset.

RQ3: *Does the number of detected change points affect the accuracy*

41

Figure 10: Matrix displaying $p-values$ of the Diebold-Mariano Test used for a pair-wise comparison of selection of the best performing models with the baseline approaches using both Hoeffding Trees and Deep learning models for the electricity load dataset.

of the forecast?

The PELT settings had a significant effect on the accuracy of the forecast. For the natural gas consumption dataset the results indicated that

fewer change points were preferable, as shown in Tables 5, 6, 7. More change points captured short-term changes in the data, often leading to overfitting of the model and increasing error. The *Low* PELT setting yielded a lower forecast error for all collection selection schemes compared to the baselines, which was not the case for other setups. The *High* PELT setting produced the highest forecast error for all approaches. This effect can be observed in Figures 14a to 14d. The figures provide a comparison of Pure Change Point-Divided Model Collection (HT-PCPDMC) approach with *Low* PELT setting with both Mixed Change Point-Divided Model Collection approaches (HT-MCPDMC-WA and HT-MCPDMC-SW) but with *High* PELT settings for the last forecasted year 2020. Figures 14a and 14b show that both HT-MCPDMC approaches with a higher number of change points have a bias towards forecasting values higher than the ground truth compared to the HT-PCPDMC approach together with a higher error variance, especially during the period from the end of March to the first half of April. To see the effect more clearly, Figure 14d and Table 2 show the aggregated errors by month; we may see that the median of the Symmetric Absolute Percentage Error in the majority of months is lower for the HT-PCPDMC approach compared to both HT-MCPDMC approaches, and the biggest difference occurred during the mentioned period. A detailed view of this period is shown in Figure 14c. We may see that the underperformance of the HT-MCPDMC approaches is mainly caused by the model performance after April 7th as the models employing the HT-MCPDMC approach with a higher number of

change points adapt to the slowly decreasing trend slower than the model based on the HT-PCPDMC approach with a lower number of change points.

The results for the electritiy load data shows that the lower number of change points were also preferable to a higher number. However, the results for *Low* and *Medium* The PELT settings show that for these settings, the results are similar, but *High* number of change points used caused an increase in error. Fig. 15a shows that models using *High* number of change points tend to overestimate the regular electricity load levels and if there is a peak in the load. We can also see in Figures 15b and 15d, and also Table 3, that during the summer months the error levels are elevated of models utilizing *High* number of change points compared to other models.

Deep learning models did not benefit from higher number of change points either and we can see in experiments using both experiments that single deep learning model outperformed every other deep learning approach, the effect is more pronounced in the natural gas consumption datasets as in some cases the deep learning models using PCPDMC model selection approach even failed to converge (see Fig. 11), thus the model switching approaches are suitable for the Hoeffding Tree-based models however not for the deep learning models, but we can see in e.g. Fig. 15c that even the single deep learning model (FCN-SMCA) without using any switching is not able to compare to the Hoeffding Tree-based models because the adaptability to changing levels of electricity consumption is lower.

Table 2: Symmetric Absolute Percentage Error (SAPE) medians of the selected forecasting approaches aggregated by months for 2020 of the natural gas consumption dataset. The lowest median error for the month is in bold.

Month/Approach	HT-PCPDMC (Low)	HT-MCPDMC-WA (High)	HT-MCPDMC-SW (High)	FCN-SMCA (M)
January	**5.53**	10.14	8.61	7.62
February	**6.10**	6.96	7.27	8.31
March	**7.78**	14.29	14.22	8.97
April	**14.74**	30.66	25.76	16.49
May	**14.69**	15.27	15.27	17.60
June	6.91	6.91	**7.53**	9.23
July	**8.20**	8.34	8.34	15.95
August	**7.33**	7.99	7.99	13.40
September	9.98	9.42	**8.42**	11.05
October	10.55	**9.57**	10.27	9.61
November	8.20	7.98	8.79	**7.85**
December	**5.66**	8.14	7.48	6.82

Table 3: Symmetric Absolute Percentage Error (SAPE) medians of the selected forecasting approaches aggregated by months for 2020 of the electricity load dataset. The lowest median error for the month is in bold.

Month/Approach	HT-PCPDMC (Low)	HT-MCPDMC-WA (High)	HT-MCPDMC-SW (High)	FCN-SMCA (S)
January	4.01	4.43	**3.37**	8.91
February	5.39	4.96	**3.48**	7.10
March	6.32	4.90	**4.32**	4.88
April	**6.70**	8.47	8.03	7.04
May	6.88	7.52	7.52	**6.14**
June	**6.57**	8.54	8.54	7.67
July	5.98	5.09	**4.54**	6.29
August	4.88	4.85	**4.66**	5.92
September	**5.15**	5.47	5.47	7.55
October	6.16	**5.68**	5.68	7.16
November	**4.59**	5.99	5.57	8.20
December	**6.83**	9.75	8.64	7.40

Table 4: Summarized results of experiments conducted over natural gas consumption dataset using the baseline approaches. The MAE, MSE and SMAPE measures were computed for the forecasted data from January 1, 2014, to December 31, 2020. SMCA = Single Model Collection Approach; QDMDC = Quarter-Divided Model Collection. Approaches prefixed with HT use Hoeffding Trees and FCN, LSTM or GRU prefixed ones use deep learning models containing Fully Connected, Long Short Term Memory or Gated Recurrent Units layers.

Approach	MAE	MSE	SMAPE
HT-SMCA	1.170e+04	3.143e+08	**12.94**
HT-QDMDC	1.179e+04	3.109e+08	13.14
FCN-SMCA (S)	1.588e+04	5.866e+08	16.70
FCN-SMCA (M)	1.548e+04	5.398e+08	16.49
FCN-SMCA (L)	1.655e+04	6.052e+08	17.09
GRU-SMCA (S)	1.936e+04	8.117e+08	20.68
GRU-SMCA (M)	1.919e+04	7.970e+08	20.26
GRU-SMCA (L)	1.798e+04	7.080e+08	19.03
LSTM-SMCA (S)	1.948e+04	7.924e+08	21.71
LSTM-SMCA (M)	1.915e+04	8.085e+08	21.65
LSTM-SMCA (L)	1.879e+04	7.557e+08	20.58
FCN-QDMDC (S)	1.983e+04	8.698e+08	21.61
FCN-QDMDC (M)	1.938e+04	8.493e+08	21.08
FCN-QDMDC (L)	2.100e+04	9.381e+08	25.77
GRU-QDMDC (S)	2.605e+04	1.497e+09	30.06
GRU-QDMDC (M)	2.306e+04	1.109e+09	27.50
GRU-QDMDC (L)	2.352e+04	1.121e+09	31.46
LSTM-QDMDC (S)	2.612e+04	1.401e+09	31.10
LSTM-QDMDC (M)	3.301e+04	3.058e+09	34.81
LSTM-QDMDC (L)	3.582e+04	3.114e+09	50.19

Table 5: Summarized results of experiments conducted over natural gas consumption dataset using the proposed approaches for the *Low* number of change points. The MAE, MSE and SMAPE measures were computed for the forecasted data from January 1, 2014, to December 31, 2020. PCPDMC = Pure Change Point-Divided Model Collection; MCPDMC = Mixed Change Point-Divided Model Collection, WA and SW suffix distinct between WAVG and SWITCH procedures. Approaches prefixed with HT use Hoeffding Trees and FCN, LSTM or GRU prefixed ones use deep learning models containing Fully Connected, Long Short Term Memory or Gated Recurrent Units layers.

Approach	MAE	MSE	SMAPE
HT-PCPDMC	1.110e+04	2.783e+08	12.32
HT-MCPDMC-WA	1.108e+04	2.770e+08	**12.29**
HT-MCPDMC-SW	1.124e+04	2.856e+08	12.48
FCN-PCPDMC (S)	1.902e+04	7.643e+08	21.41
FCN-PCPDMC (M)	1.750e+04	6.457e+08	20.34
FCN-PCPDMC (L)	1.780e+04	6.650e+08	21.49
GRU-PCPDMC (S)	2.325e+04	1.115e+09	28.11
GRU-PCPDMC (M)	2.185e+04	1.003e+09	24.89
GRU-PCPDMC (L)	2.107e+04	9.241e+08	24.55
LSTM-PCPDMC (S)	2.786e+04	1.781e+09	45.80
LSTM-PCPDMC (M)	2.533e+04	1.081e+09	44.68
LSTM-PCPDMC (L)	2.143e+04	9.764e+08	24.23

Table 6: Summarized results of experiments conducted over natural gas consumption dataset using the proposed approaches for the *Medium* number of change points. The MAE, MSE and SMAPE measures were computed for the forecasted data from January 1, 2014, to December 31, 2020. PCPDMC = Pure Change Point-Divided Model Collection; MCPDMC = Mixed Change Point-Divided Model Collection, WA and SW suffix distinct between WAVG and SWITCH procedures. Approaches prefixed with HT use Hoeffding Trees and FCN, LSTM or GRU prefixed ones use deep learning models containing Fully Connected, Long Short Term Memory or Gated Recurrent Units layers.

Approach	MAE	MSE	SMAPE
HT-PCPDMC	1.181e+04	3.211e+08	**12.60**
HT-MCPDMC-WA	1.196e+04	3.233e+08	12.78
HT-MCPDMC-SW	1.189e+04	3.223e+08	12.81
FCN-PCPDMC (S)	2.090e+04	9.454e+08	22.43
FCN-PCPDMC (M)	2.030e+04	8.837e+08	22.17
FCN-PCPDMC (L)	2.010e+04	8.929e+08	22.08
GRU-PCPDMC (S)	2.774e+04	1.331e+09	45.60
GRU-PCPDMC (M)	2.299e+04	1.105e+09	26.51
GRU-PCPDMC (L)	2.270e+04	1.091e+09	26.54
LSTM-PCPDMC (S)	4.856e+04	8.031e+09	51.02
LSTM-PCPDMC (M)	3.275e+04	3.208e+09	45.85
LSTM-PCPDMC (L)	2.187e+04	1.053e+09	23.87

Table 7: Summarized results of experiments conducted over natural gas consumption dataset using the proposed approaches for the *High* number of change points. The MAE, MSE and SMAPE measures were computed for the forecasted data from January 1, 2014, to December 31, 2020. PCPDMC = Pure Change Point-Divided Model Collection; MCPDMC = Mixed Change Point-Divided Model Collection, WA and SW suffix distinct between WAVG and SWITCH procedures. Approaches prefixed with HT use Hoeffding Trees and FCN, LSTM or GRU prefixed ones use deep learning models containing Fully Connected, Long Short Term Memory or Gated Recurrent Units layers.

Approach	MAE	MSE	SMAPE
HT-PCPDMC	1.265e+04	3.925e+08	**13.09**
HT-MCPDMC-WA	1.557e+04	6.368e+08	15.12
HT-MCPDMC-SW	1.483e+04	5.711e+08	14.72
FCN-PCPDMC (S)	2.100e+04	9.688e+08	22.28
FCN-PCPDMC (M)	1.975e+04	8.508e+08	21.05
FCN-PCPDMC (L)	1.977e+04	8.391e+08	22.16
GRU-PCPDMC (S)	2.450e+04	1.265e+09	29.71
GRU-PCPDMC (M)	2.264e+04	1.029e+09	27.60
GRU-PCPDMC (L)	2.173e+04	9.828e+08	25.01
LSTM-PCPDMC (S)	2.498e+04	1.387e+09	29.19
LSTM-PCPDMC (M)	3.183e+04	3.135e+09	45.33
LSTM-PCPDMC (L)	2.097e+04	9.366e+08	23.11

Table 8: Summarized results of experiments conducted over electricity load dataset using the baseline approaches. The MAE, MSE and SMAPE measures were computed for the forecasted data from January 1, 2014, to December 31, 2020. SMCA = Single Model Collection Approach; QDMDC = Quarter-Divided Model Collection. Approaches prefixed with HT use Hoeffding Trees and FCN, LSTM or GRU prefixed ones use deep learning models containing Fully Connected, Long Short Term Memory or Gated Recurrent Units layers.

Approach	MAE	MSE	SMAPE
HT-SMCA	507.2	5.053e+05	**6.84**
HT-QDMDC	537.6	5.480e+05	7.25
FCN-SMCA (S)	573.5	5.571e+05	7.73
FCN-SMCA (M)	592.1	5.906e+05	7.97
FCN-SMCA (L)	580.7	5.764e+05	7.82
GRU-SMCA (S)	673.2	7.455e+05	9.10
GRU-SMCA (M)	642.2	6.975e+05	8.68
GRU-SMCA (L)	636.1	6.873e+05	8.60
LSTM-SMCA (S)	683.4	7.664e+05	9.26
LSTM-SMCA (M)	627.6	6.698e+05	8.49
LSTM-SMCA (L)	646.3	7.036e+05	8.74
FCN-QDMDC (S)	659.5	6.967e+05	8.87
FCN-QDMDC (M)	646.1	6.733e+05	8.68
FCN-QDMDC (L)	646.5	6.749e+05	8.69
GRU-QDMDC (S)	690.8	7.757e+05	9.34
GRU-QDMDC (M)	668.0	7.338e+05	9.05
GRU-QDMDC (L)	672.0	7.493e+05	9.09
LSTM-QDMDC (S)	842.1	1.428e+06	11.45
LSTM-QDMDC (M)	673.8	7.459e+05	9.11
LSTM-QDMDC (L)	665.4	7.370e+05	9.00

Table 9: Summarized results of experiments conducted over electricity load dataset using the proposed approaches for the *Low* number of change points. The MAE, MSE and SMAPE measures were computed for the forecasted data from January 1, 2014, to December 31, 2020. PCPDMC = Pure Change Point-Divided Model Collection; MCPDMC = Mixed Change Point-Divided Model Collection, WA and SW suffix distinct between WAVG and SWITCH procedures. Approaches prefixed with HT use Hoeffding Trees and FCN, LSTM or GRU prefixed ones use deep learning models containing Fully Connected, Long Short Term Memory or Gated Recurrent Units layers.

Approach	MAE	MSE	SMAPE
HT-PCPDMC	506.5	5.020e+05	**6.83**
HT-MCPDMC-WA	522.6	5.344e+05	7.05
HT-MCPDMC-SW	526.6	5.445e+05	7.10
FCN-PCPDMC (S)	624.6	6.380e+05	8.38
FCN-PCPDMC (M)	611.6	6.251e+05	8.22
FCN-PCPDMC (L)	612.2	6.276e+05	8.23
GRU-PCPDMC (S)	649.3	7.063e+05	8.77
GRU-PCPDMC (M)	642.3	6.948e+05	8.68
GRU-PCPDMC (L)	634.0	6.783e+05	8.57
LSTM-PCPDMC (S)	648.1	7.066e+05	8.74
LSTM-PCPDMC (M)	635.2	6.838e+05	8.58
LSTM-PCPDMC (L)	633.1	6.768e+05	8.56

Table 10: Summarized results of experiments conducted over electricity load dataset using the proposed approaches for the *Medium* number of change points. The MAE, MSE and SMAPE measures were computed for the forecasted data from January 1, 2014, to December 31, 2020. PCPDMC = Pure Change Point-Divided Model Collection; MCPDMC = Mixed Change Point-Divided Model Collection, WA and SW suffix distinct between WAVG and SWITCH procedures. Approaches prefixed with HT use Hoeffding Trees and FCN, LSTM or GRU prefixed ones use deep learning models containing Fully Connected, Long Short Term Memory or Gated Recurrent Units layers.

Approach	MAE	MSE	SMAPE
HT-PCPDMC	506.2	5.001e+05	**6.83**
HT-MCPDMC-WA	532.1	5.467e+05	7.17
HT-MCPDMC-SW	533.3	5.555e+05	7.11
FCN-PCPDMC (S)	625.2	6.473e+05	8.40
FCN-PCPDMC (M)	616.2	6.368e+05	8.27
FCN-PCPDMC (L)	615.1	6.361e+05	8.26
GRU-PCPDMC (S)	649.3	7.091e+05	8.76
GRU-PCPDMC (M)	637.8	6.847e+05	8.63
GRU-PCPDMC (L)	634.7	6.865e+05	8.58
LSTM-PCPDMC (S)	662.7	7.426e+05	8.92
LSTM-PCPDMC (M)	724.7	1.002e+06	9.55
LSTM-PCPDMC (L)	625.6	6.661e+05	8.46

Table 11: Summarized results of experiments conducted over electricity load dataset using the proposed approaches for the *High* number of change points. The MAE, MSE and SMAPE measures were computed for the forecasted data from January 1, 2014, to December 31, 2020. PCPDMC = Pure Change Point-Divided Model Collection; MCPDMC = Mixed Change Point-Divided Model Collection, WA and SW suffix distinct between WAVG and SWITCH procedures. Approaches prefixed with HT use Hoeffding Trees and FCN, LSTM or GRU prefixed ones use deep learning models containing Fully Connected, Long Short Term Memory or Gated Recurrent Units layers.

Approach	MAE	MSE	SMAPE
HT-PCPDMC	517.2	5.197e+05	**6.99**
HT-MCPDMC-WA	576.0	6.168e+05	7.77
HT-MCPDMC-SW	573.1	6.278e+05	7.74
FCN-PCPDMC (S)	669.2	7.163e+05	9.01
FCN-PCPDMC (M)	650.5	6.817e+05	8.75
FCN-PCPDMC (L)	653.1	6.855e+05	8.79
GRU-PCPDMC (S)	696.5	7.957e+05	9.42
GRU-PCPDMC (M)	675.3	7.422e+05	9.12
GRU-PCPDMC (L)	667.0	7.297e+05	9.01
LSTM-PCPDMC (S)	707.6	8.227e+05	9.55
LSTM-PCPDMC (M)	674.7	7.422e+05	9.12
LSTM-PCPDMC (L)	664.6	7.252e+05	9.00
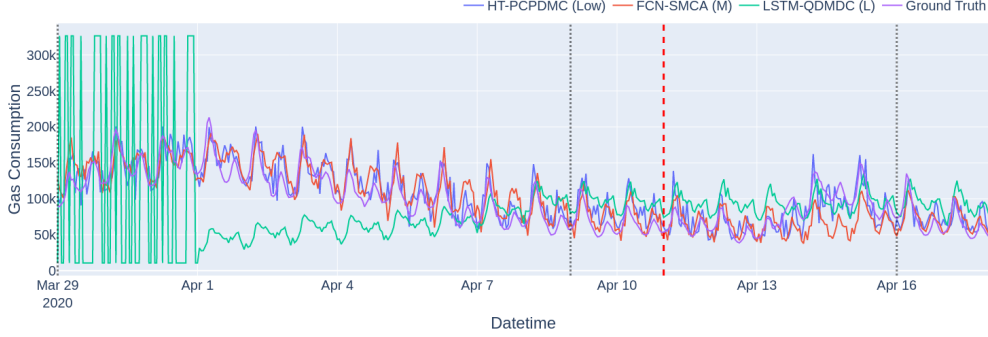
Figure 11: Example of an extreme behaviour of a deep learning model (LSTM-QDMDC (L)) which failed to coverge when there were different models used in different segments.

RQ4: Is there an advantage in employing forecast ensembling (MCPDMC-WA) or model switching (MCPDMC-SW) in proximity to change points, with the main consideration being the Symmetric Mean Absolute Percentage Error (SMAPE) metric?

The performance of the schemas for Mixed Change Point-Divided Model Collection (HT-MCPDMC-WA and HT-MCPDMC-SW) depended on the density of change points (i.e., number of segments). In case of natural gas consumption forecasting and *Low* PELT settings, the WAVG procedure performed similarly to the HT-PCPDMC approach and slightly reduced the forecast error for all metrics (Table 9 and Figure 12a). However, with more change points, the error reduction was only observed in the first few years of the data, as shown in Figs. 12b and 12c. After 2015, the HT-PCPDMC approach had sub 13% SMAPE while both WAVG and SWITCH procedures had SMAPE over 15 % and failed to beat the baselines. The HT-PCPDMC approach also showed a decreasing trend in SMAPE during incremental learning, similar to

54

the baseline models, but with a lower SMAPE for all PELT settings. Therefore, the HT-PCPDMC approach was more suitable for CL tasks than the HT-MCPDMC approach, as it was less sensitive to the number of detected change points and achieved better results with a simpler procedure. Furthermore, the HT-PCPDMC approach had the lowest final SMAPE (year 2020) compared to all other approaches for each PELT setting, as seen in Figs. 12a to 12c.

We can see the same effect in the electricity load forecasting as in neither PELT setting the WAVG or SWITCH procedures provided lower forecasting error compared to the HT-PCPDMC approach (see Tables 9) and 11 or Figures 13a and 13c.

7. Conclusions and future works

This article proposed and evaluated two novel approaches to forecasting natural gas consumption based on change point detection and model aggregation. We designed a general methodology for multistep forecasting of multivariate time series with CL capabilities using data stream processing. The methodology consists of several steps, such as dataset preprocessing, change point detection, model collection setup, model collection selection schema, and model training. We compared the performance of the pure and mixed change point divided model collection approaches with two baseline approaches that do not use change point detection and also deep learning models wihh LSTM, GRU or Dense layers used. Using two real-world datasets from the energy domain. We used three metrics (MSE, MAE and SMAPE), Diebold-Mariano test and three settings of the PELT algorithm to

55

measure the forecasting error and the effect of the number of change points. Our experiments showed that the pure change point divided model collection approach outperformed all other approaches by all metrics and for all settings. Deep learning models did not fare as well, showing higher errors across various architectures, with little benefit from employing multiple models or change point selection. These findings suggest that Hoeffding Tree-based models, especially those employing the HT-PCPDMC scheme, are more effective in forecasting natural gas consumption or electricity load compared to deep learning approaches. We also found that fewer change points resulted in a lower forecasting error and that the pure approach was more robust and suitable for CL tasks than the mixed approach. Our results demonstrate the potential for using change point detection and model aggregation to forecast natural gas consumption in a dynamic and complex environment.

Using the Hoeffding Tree Regressor for multistep time series forecasting in data streams offers several advantages, particularly in scenarios where data arrives continuously and in large volumes. However, certain factors must be considered before applying this approach. First, if the dataset lacks sufficient data or exhibits significant data shifts, it may be more appropriate to utilize offline learning techniques. Offline learning allows the utilization of the entire dataset, enabling more comprehensive model training and evaluation. In contrast, in data stream scenarios with limited data availability, the Hoeffding Tree regressor may not perform optimally due to insufficient data for model training and adaptation. Moreover, the employment of the Pruned Exact Linear Time (PELT) algorithm for change point detection in time series data streams is beneficial when the time series contains data shifts or

structural changes. In cases where the time series exhibits stability and does not undergo significant shifts, the use of PELT may not be necessary. However, when data shifts are present, employing PELT can improve forecasting accuracy by identifying and adapting to these changes in the data distribution. Comparatively, utilizing the Hoeffding Tree regression with PELT change point detection often yields superior results compared to deep learning approaches in scenarios characterized by data shifts. The adaptability of the Hoeffding Tree Regressor to evolving data distributions, coupled with the ability of PELT to detect and respond to changes, contributes to its effectiveness in forecasting time series data streams.

We believe that this is an important and promising direction for future research, as it could improve the accuracy and efficiency of forecasting systems and enable them to adapt to changing conditions. Some possible future work includes exploring different types of models and selection schemes, applying the methodology to other domains and datasets, and incorporating uncertainty estimation and anomaly detection into the forecasting pipeline.
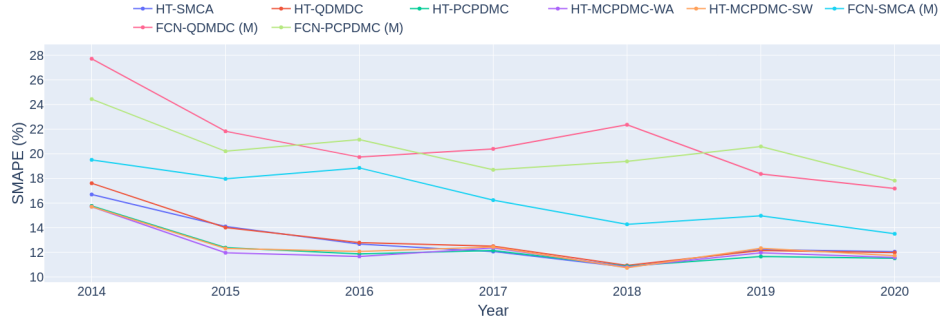
(a) SMAPE yearly aggregated progress during test-then-train experiment protocol for PELT Settings = *Low*.
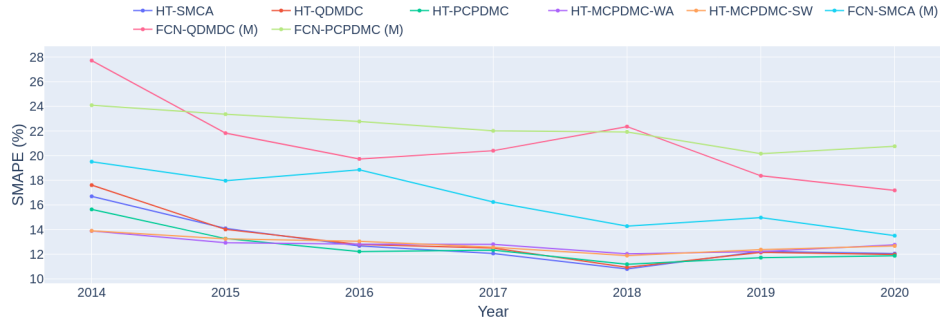
(b) SMAPE yearly aggregated progress during test-then-train experiment protocol for PELT Settings = *Medium*.
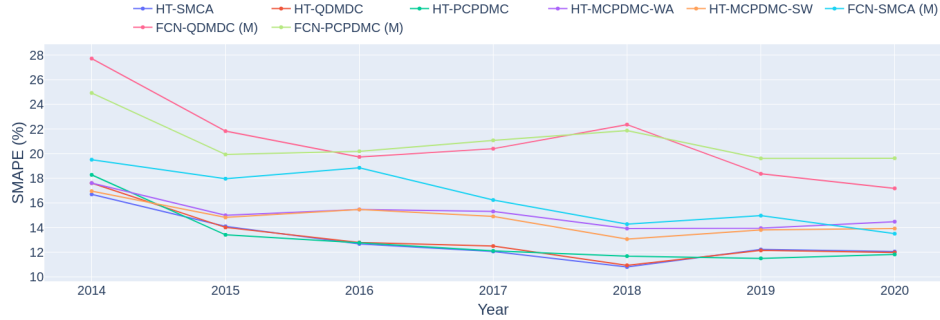
(c) SMAPE yearly aggregated progress during test-then-train experiment protocol for PELT Settings = *High*.
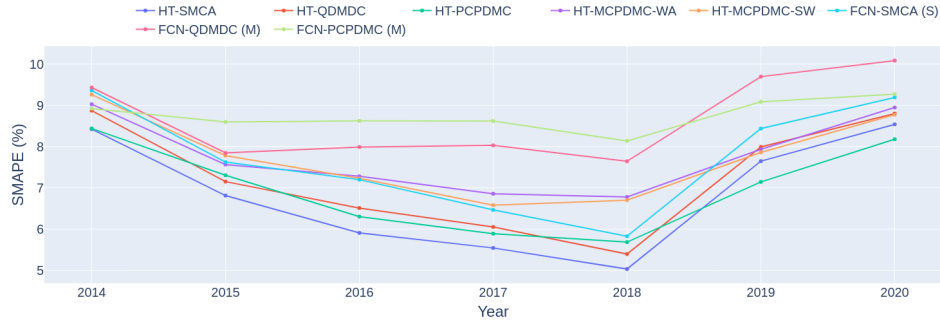
Figure 12: Progress of yearly aggregated SMAPE for the conducted experiments using natural gas consumption dataset, divided by the change point detection algorithm setting, proposed models, and baseline comparison, is included for each setting. The SMAPE error measure was computed for the forecasted data from January 1, 2014, to December 31, 2020. SMCA = Single Model Collection Approach; QDMDC = Quarter-Divided Model Collection; PCPDMC = Pure Change Point-Divided Model Collection; MCPDMC = Mixed Change Point-Divided Model Collection, WA and SW suffix distinct between WAVG and SWITCH procedures; FCN = Fully Connected Network, (S), (M), (L) suffix means which of the Small, Medium or Large size configuration of the network was used.

(a) SMAPE yearly aggregated progress during test-then-train experiment protocol for PELT Settings = *Low*.
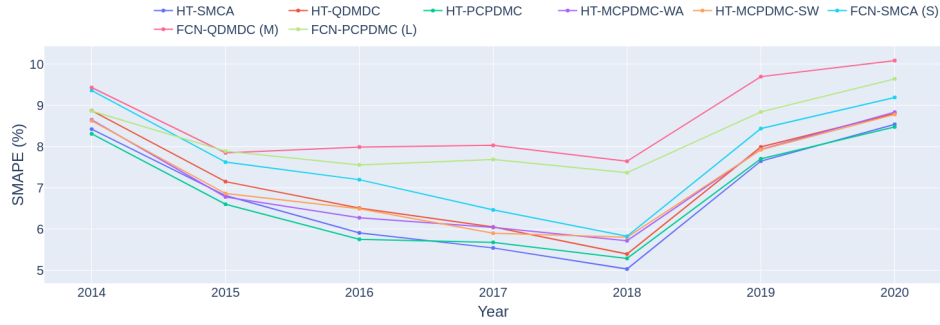
(b) SMAPE yearly aggregated progress during test-then-train experiment protocol for PELT Settings = *Medium*.

(c) SMAPE yearly aggregated progress during test-then-train experiment protocol for PELT Settings = *High*.

Figure 13: Progress of yearly aggregated SMAPE for the conducted experiments using electricity load dataset, divided by the change point detection algorithm setting, proposed models, and baseline comparison, is included for each setting. The SMAPE error measure was computed for the forecasted data from January 1, 2014, to December 31, 2020. SMCA = Single Model Collection Approach; QDMDC = Quarter-Divided Model Collection; PCPDMC = Pure Change Point-Divided Model Collection; MCPDMC = Mixed Change Point-Divided Model Collection, WA and SW suffix distinct between WAVG and SWITCH procedures; FCN = Fully Connected Network, (S), (M), (L) suffix means which of the Small, Medium or Large size configuration of the network was used.
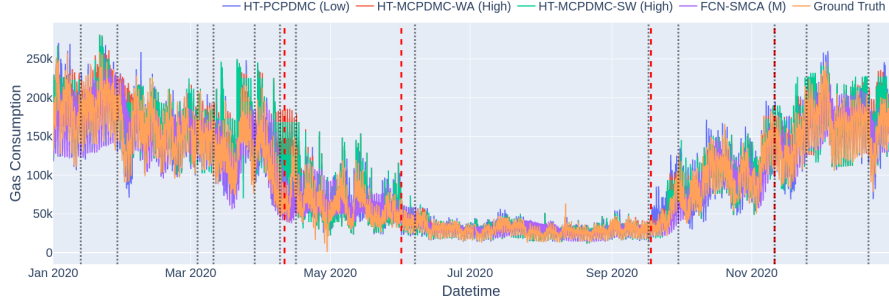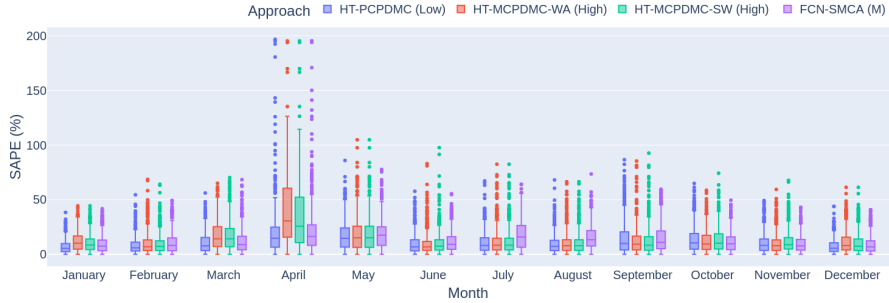
(a) Ground truth natural gas consumption compared with forecasted values using selected approaches for 2020.
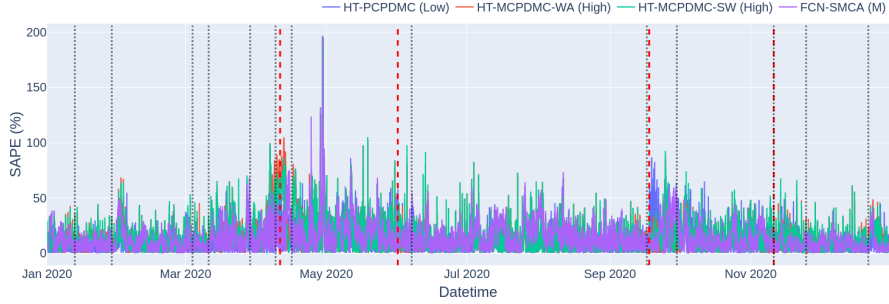
(b) Symmetric Absolute Percentage Error (SAPE) plot of the selected forecasting approaches for 2020.
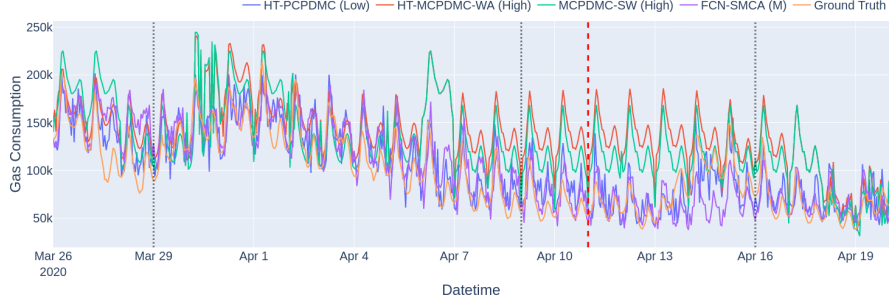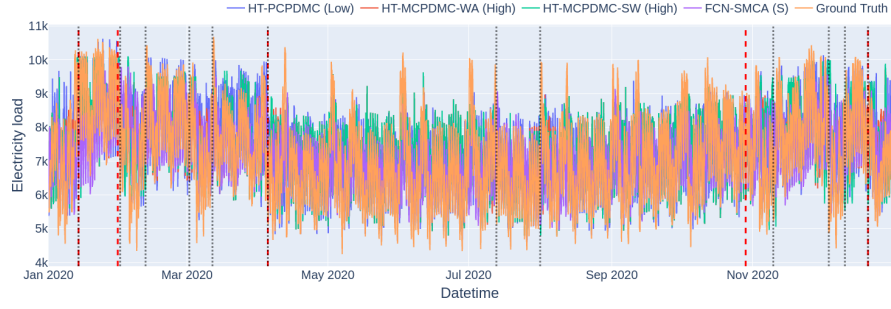

(c) Ground truth natural gas consumption compared with forecasted values using selected approaches for the selected part of the year 2020 in which a multiple change points occur and significant forecasting accuracy difference among the methods is present.
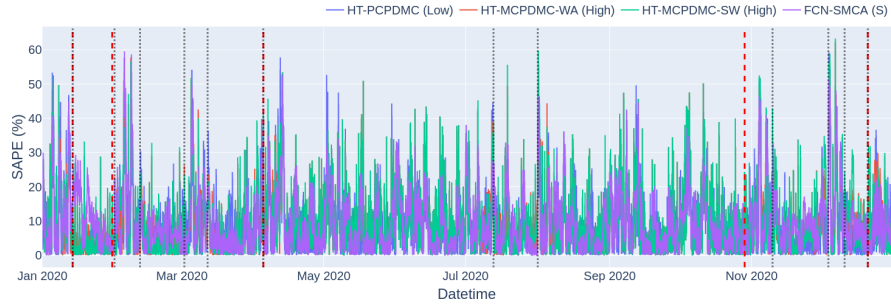

(d) Symmetric Absolute Percentage Error (SAPE) of the selected forecasting approaches aggregated by months for 2020.
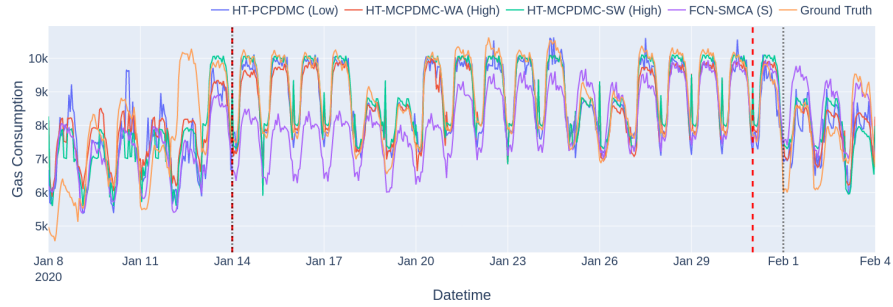
60

Figure 14: Detailed look at results of PCPDMC (*Low* PELT settings), MCPDMC-WA, MCPDMC-SW (*High* PELT settings for both) and FCN-SMCA (M) approaches for the 2020 natural gas consumption data. The red dashed and black dotted vertical lines present in subfigures (a) to (c) represent the change point locations detected with *Low*, respectively *High*, PELT algorithm settings.

(a) Ground truth electricity load compared with forecasted values using selected approaches for 2020.

(b) Symmetric Absolute Percentage Error (SAPE) plot of the selected forecasting approaches for 2020.

(c) Ground truth electricity load compared with forecasted values using selected approaches for the selected part of the year 2020 in which a multiple change points occur and significant forecasting accuracy difference among the methods is present.
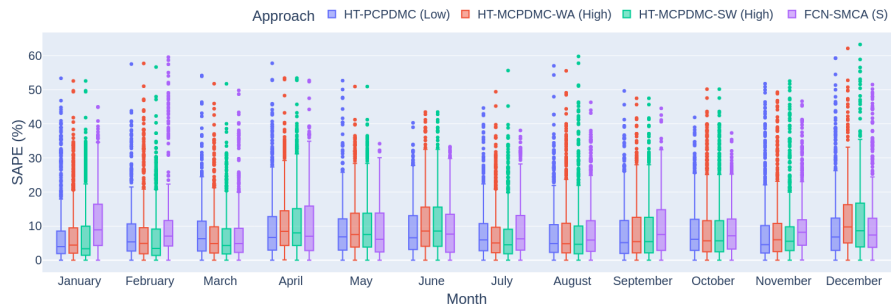
(d) Symmetric Absolute Percentage Error (SAPE) of the selected forecasting approaches aggregated by months for 2020.

Figure 15: Detailed look at results of PCPDMC (*Low* PELT settings), MCPDMC-WA, MCPDMC-SW (*High* PELT settings for both) and FCN-SMCA (S) approaches for the 2020 electricity load data. The red dashed and black dotted vertical lines present in subfigures (a) to (c) represent the change point locations detected with *Low*, respectively *High*, PELT algorithm settings.

Acknowledgment

References

S. E. Masten, K. J. Crocker, Efficient adaptation in long-term contracts: Take-or-pay provisions for natural gas, The American Economic Review 75 (5) (1985) 1083–1093.

A. Creti, B. Villeneuve, et al., Long-term contracts and take-or-pay clauses in natural gas markets, Energy Studies Review 13 (1) (2004) 75–94.

J. M. Medina, G. A. McKenzie, B. M. Daniel, Take or litigate: Enforcing the plain meaning of the take-or-pay clause in natural gas contracts, Ark. L. Rev. 40 (1986) 185.

P. Balestra, M. Nerlove, Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas, Econometrica 34 (3) (1966) 585–612.
URL http://www.jstor.org/stable/1909771

J. Vondráček, E. Pelikán, O. Konár, J. Čermáková, K. Eben, M. Malý, M. Brabec, A statistical model for the estimation of natural gas consumption, Applied Energy 85 (5) (2008) 362 – 370. doi:https://doi.org/10.1016/j.apenergy.2007.07.004.

URL   http://www.sciencedirect.com/science/article/pii/
S0306261907001183

J. Suykens, P. Lemmerling, W. Favoreel, B. De Moor, M. Crepel, P. Briol,
Modelling the belgian gas consumption using neural networks, Neural
processing letters 4 (3) (1996) 157–166.

A. Khotanzad, H. Elragal, T. . Lu, Combination of artificial neural-network
forecasters for prediction of natural gas consumption, IEEE Transactions
on Neural Networks 11 (2) (2000) 464–473.

N. H. Viet, J. Mańdziuk, Neural and fuzzy neural networks in prediction of
natural gas consumption, Neural Parallel & Scientific Comp. 13 (2005)
265–286.

B. Soldo, P. Potočnik, G. Šimunović, T. Šarić, E. Govekar, Improv-
ing the residential natural gas consumption forecasting models by
using solar radiation, Energy and Buildings 69 (2014) 498 – 506.
doi:https://doi.org/10.1016/j.enbuild.2013.11.032.
URL   http://www.sciencedirect.com/science/article/pii/
S0378778813007299

F. Taşpınar, N. Çelebi, N. Tutkun, Forecasting of daily natural gas
consumption on regional basis in turkey using various com-
putational methods, Energy and Buildings 56 (2013) 23 – 31.
doi:https://doi.org/10.1016/j.enbuild.2012.10.023.
URL   http://www.sciencedirect.com/science/article/pii/
S0378778812005324

F. E. Boran, Forecasting natural gas consumption in turkey using grey prediction, Energy Sources, Part B: Economics, Planning, and Policy 10 (2) (2015) 208–213. `arXiv:https://doi.org/10.1080/15567249.2014.893040`, `doi:10.1080/15567249.2014.893040`.
URL `https://doi.org/10.1080/15567249.2014.893040`

J. Szoplik, Forecasting of natural gas consumption with artificial neural networks, Energy 85 (2015) 208 – 220. `doi:https://doi.org/10.1016/j.energy.2015.03.084`.
URL `http://www.sciencedirect.com/science/article/pii/S036054421500393X`

A. Rahman, A. D. Smith, Predicting fuel consumption for commercial buildings with machine learning algorithms, Energy and Buildings 152 (2017) 341 – 358. `doi:https://doi.org/10.1016/j.enbuild.2017.07.017`.

H. Su, E. Zio, J. Zhang, M. Xu, X. Li, Z. Zhang, A hybrid hourly natural gas demand forecasting method based on the integration of wavelet transform and enhanced deep-rnn model, Energy 178 (2019) 585 – 597. `doi:https://doi.org/10.1016/j.energy.2019.04.167`.

Y. Bai, C. Li, Daily natural gas consumption forecasting based on a structure-calibrated support vector regression approach, Energy and Buildings 127 (2016) 571 – 579. `doi:https://doi.org/10.1016/j.enbuild.2016.06.020`.
URL `http://www.sciencedirect.com/science/article/pii/S0378778816305096`

B. Soldo, Forecasting natural gas consumption, Applied Energy 92 (2012) 26 – 37. `doi:https://doi.org/10.1016/j.apenergy.2011.11.003`.

J. G. Tamba, S. N. Essiane, E. F. Sapnken, F. D. Koffi, J. L. Nsouandélé, B. Soldo, D. Njomo, Forecasting Natural Gas: A Literature Survey, International Journal of Energy Economics and Policy 8 (3) (2018) 216–249.
URL `https://ideas.repec.org/a/eco/journ2/2018-03-28.html`

Z. Chen, B. Liu, R. Brachman, P. Stone, F. Rossi, Lifelong Machine Learning, 2nd Edition, Morgan & Claypool Publishers, 2018.

R. M. French, Catastrophic forgetting in connectionist networks, Trends in Cognitive Sciences 3 (4) (1999) 128–135. `doi:https://doi.org/10.1016/S1364-6613(99)01294-2`.
URL `https://www.sciencedirect.com/science/article/pii/S1364661399012942`

J. Knoblauch, H. Husain, T. Diethe, Optimal continual learning has perfect memory and is np-hard, in: ICML 2020, 2020.
URL `https://www.amazon.science/publications/optimal-continual-learning-has-perfect-memory-and-is-np-hard`

G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, Neural Networks 113 (2019) 54–71. `doi:https://doi.org/10.1016/j.neunet.2019.01.012`.
URL `https://www.sciencedirect.com/science/article/pii/S0893608019300231`

A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. S. Torr, M. Ranzato, Continual learning with tiny episodic memories, CoRR abs/1902.10486 (2019). `arXiv:1902.10486`.
URL `http://arxiv.org/abs/1902.10486`

R. Aljundi, E. Belilovsky, T. Tuytelaars, L. Charlin, M. Caccia, M. Lin, L. Page-Caccia, Online continual learning with maximal interfered retrieval, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019.
URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/15825aee15eb335cc13f9b559f166ee8-Paper.pdf`

L. Caccia, R. Aljundi, N. Asadi, T. Tuytelaars, J. Pineau, E. Belilovsky, New insights on reducing abrupt representation change in online continual learning, in: International Conference on Learning Representations, 2022.
URL `https://openreview.net/forum?id=N8MaByOzUfb`

Y. Zhang, B. Pfahringer, E. Frank, A. Bifet, N. J. S. Lim, Y. Jia, A simple but strong baseline for online continual learning: Repeated augmented rehearsal, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, Vol. 35, Curran Associates, Inc., 2022, pp. 14771–14783.
URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/5ebbbac62b968254093023f1c95015d3-Paper-Conference.pdf`

P. Buzzega, M. Boschini, A. Porrello, D. Abati, S. CALDERARA, Dark ex-

perience for general continual learning: a strong, simple baseline, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 15920–15930.
URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/b704ea2c39778f07c617f6b7ce480e9e-Paper.pdf`

M. Boschini, L. Bonicelli, P. Buzzega, A. Porrello, S. Calderara, Class-incremental continual learning into the extended der-verse, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (5) (2023) 5497–5512. `doi:10.1109/TPAMI.2022.3206549`.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks, Proceedings of the National Academy of Sciences 114 (13) (2017) 3521–3526. `arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1611835114`, `doi:10.1073/pnas.1611835114`.
URL `https://www.pnas.org/doi/abs/10.1073/pnas.1611835114`

F. Zenke, B. Poole, S. Ganguli, Continual learning through synaptic intelligence, Proceedings of machine learning research 70 (2017) 3987–3995.

A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, R. Hadsell, Progressive neural networks, CoRR abs/1606.04671 (2016). `arXiv:1606.04671`.
URL `http://arxiv.org/abs/1606.04671`

B. Ermis, G. Zappella, M. Wistuba, A. Rawal, C. Archambeau, Memory efficient continual learning with transformers, in: A. H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural Information Processing Systems, 2022.
URL https://openreview.net/forum?id=U07d1Y-x2E

V. Gupta, J. Narwariya, P. Malhotra, L. Vig, G. Shroff, Continual learning for multivariate time series tasks with variable input dimensions, in: 2021 IEEE International Conference on Data Mining (ICDM), 2021, pp. 161–170. doi:10.1109/ICDM51629.2021.00026.

G. G. González, P. Casas, A. Fernández, G. Gómez, Steps towards continual learning in multivariate time-series anomaly detection using variational autoencoders, in: Proceedings of the 22nd ACM Internet Measurement Conference, IMC '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 774–775. doi:10.1145/3517745.3563033.
URL https://doi.org/10.1145/3517745.3563033

A. Cossu, A. Carta, V. Lomonaco, D. Bacciu, Continual learning for recurrent neural networks: An empirical evaluation, Neural Networks 143 (2021) 607–627. doi:https://doi.org/10.1016/j.neunet.2021.07.021.
URL https://www.sciencedirect.com/science/article/pii/S0893608021002847

L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees, 1984.

J. R. Quinlan, Induction of decision trees, Machine Learning 1 (1) (1986)

81–106. `doi:10.1007/BF00116251`.

URL `https://doi.org/10.1007/BF00116251`

J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

P. Domingos, G. Hulten, Mining high-speed data streams, in: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '00, Association for Computing Machinery, New York, NY, USA, 2000, p. 71–80. `doi:10.1145/347090.347107`.

URL `https://doi.org/10.1145/347090.347107`

C. Manapragada, G. I. Webb, M. Salehi, Extremely fast decision tree, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1953–1962. `doi: 10.1145/3219819.3220005`.

URL `https://doi.org/10.1145/3219819.3220005`

H. Abdulsalam, D. Skillicorn, P. Martin, Streaming random forests, 2007, pp. 225 – 232. `doi:10.1109/IDEAS.2007.4318108`.

A. Bifet, E. Frank, G. Holmes, B. Pfahringer, Ensembles of restricted hoeffding trees, ACM Trans. Intell. Syst. Technol. 3 (2) (feb 2012). `doi:10.1145/2089094.2089106`.

URL `https://doi.org/10.1145/2089094.2089106`

P. Cal, M. Woźniak, Parallel hoeffding decision tree for streaming data, in: S. Omatu, J. Neves, J. M. C. Rodriguez, J. F. Paz Santana, S. R. Gon-

zalez (Eds.), Distributed Computing and Artificial Intelligence, Springer International Publishing, Cham, 2013, pp. 27–35.

A. Lung-Yut-Fong, C. Lévy-Leduc, O. Cappé, Homogeneity and change-point detection tests for multivariate data using rank statistics, Journal de la société française de statistique 156 (4) (2015) 133–162.

C. Truong, L. Oudre, N. Vayatis, A review of change point detection methods, CoRR abs/1801.00718 (2018). `arXiv:1801.00718`.
URL `http://arxiv.org/abs/1801.00718`

K. Haynes, I. Eckley, P. Fearnhead, Efficient penalty search for multiple changepoint problems (12 2014).

S. Aminikhanghahi, D. J. Cook, A survey of methods for time series change point detection, Knowledge and information systems 51 (2) (2017) 339–367.

R. Killick, P. Fearnhead, I. A. Eckley, Optimal detection of changepoints with a linear computational cost, Journal of the American Statistical Association 107 (500) (2012) 1590–1598. `doi:"10.1080/01621459.2012.737745"`.
URL `https://doi.org/10.1080%2F01621459.2012.737745`

B. Jackson, J. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, T. T. Tsai, An algorithm for optimal partitioning of data on an interval, IEEE Signal Processing Letters 12 (2) (2005) 105–108. `doi:10.1109/LSP.2001.838216`.

A. J. Scott, M. Knott, A cluster analysis method for grouping means in the analysis of variance, Biometrics 30 (3) (1974) 507–512.
URL `http://www.jstor.org/stable/2529204`

R. P. Adams, D. J. C. Mackay, Bayesian online changepoint detection, arXiv: Machine Learning (2007).

IEEE, 2021. `doi:"10.1109/bigdata52589.2021.9671962"`, [link].
URL `"https://doi.org/10.1109%2Fbigdata52589.2021.9671962"`

M. Tong, F. Qin, J. Dong, Natural gas consumption forecasting using an optimized grey bernoulli model: The case of the world's top three natural gas consumers, Engineering Applications of Artificial Intelligence 122 (2023) 106005. `doi:https://doi.org/10.1016/j.engappai.2023.106005`.
URL `https://www.sciencedirect.com/science/article/pii/S0952197623001896`

A. Hussain, J. A. Memon, M. Murshed, U. H. Md Shabbir Alam, Muhammad Rahman, A time series forecasting analysis of overall and sector-based natural gas demand: a developing south asian economy case, Environmental Science and Pollution Research 29 (2022). `doi:https://doi.org/10.1007/s11356-022-20861-3`.

S.-Y. Shin, H.-G. Woo, Energy consumption forecasting in korea using machine learning algorithms, Energies 15 (13) (2022). `doi:10.3390/en15134880`.
URL `https://www.mdpi.com/1996-1073/15/13/4880`

R. Svoboda, V. Kotik, J. Platos, Short-term natural gas consumption forecasting from long-term data collection, Energy 218 (2021) 119430. `doi:https://doi.org/10.1016/j.energy.2020.119430`.
URL `http://www.sciencedirect.com/science/article/pii/S0360544220325378`

P. Domingos, G. Hulten, Mining high-speed data streams, in: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '00, Association for Computing Machinery, New York, NY, USA, 2000, p. 71–80. `doi:10.1145/347090.347107`.
URL `https://doi.org/10.1145/347090.347107`

S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (8) (1997) 1735–1780. `arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf`, `doi:10.1162/neco.1997.9.8.1735`.
URL `https://doi.org/10.1162/neco.1997.9.8.1735`

B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, M. Woźniak, Ensemble learning for data stream analysis: A survey, Information Fusion 37 (2017) 132–156. `doi:https://doi.org/10.1016/j.inffus.2017.02.004`.
URL `https://www.sciencedirect.com/science/article/pii/S1566253516302329`

A. Bifet, R. Gavaldà, G. Holmes, B. Pfahringer, Machine Learning for Data Streams with Practical Examples in MOA, MIT Press, Cambridge, MA,

2018.

URL `https://moa.cms.waikato.ac.nz/book-html/`

F. Petropoulos, D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. Ben Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, J. Browell, C. Carnevale, J. L. Castle, P. Cirillo, M. P. Clements, C. Cordeiro, F. L. Cyrino Oliveira, S. De Baets, A. Dokumentov, J. Ellison, P. Fiszeder, P. H. Franses, D. T. Frazier, M. Gilliland, M. S. Gönül, P. Goodwin, L. Grossi, Y. Grushka-Cockayne, M. Guidolin, M. Guidolin, U. Gunter, X. Guo, R. Guseo, N. Harvey, D. F. Hendry, R. Hollyman, T. Januschowski, J. Jeon, V. R. R. Jose, Y. Kang, A. B. Koehler, S. Kolassa, N. Kourentzes, S. Leva, F. Li, K. Litsiou, S. Makridakis, G. M. Martin, A. B. Martinez, S. Meeran, T. Modis, K. Nikolopoulos, D. Önkal, A. Paccagnini, A. Panagiotelis, I. Panapakidis, J. M. Pavía, M. Pedio, D. J. Pedregal, P. Pinson, P. Ramos, D. E. Rapach, J. J. Reade, B. Rostami-Tabar, M. Rubaszek, G. Sermpinis, H. L. Shang, E. Spiliotis, A. A. Syntetos, P. D. Talagala, T. S. Talagala, L. Tashman, D. Thomakos, T. Thorarinsdottir, E. Todini, J. R. Trapero Arenas, X. Wang, R. L. Winkler, A. Yusupova, F. Ziel, Forecasting: theory and practice, International Journal of Forecasting 38 (3) (2022) 705–871. `doi:https://doi.org/10.1016/j.ijforecast.2021.11.001`.
URL `https://www.sciencedirect.com/science/article/pii/S0169207021001758`

F. Diebold, R. Mariano, Comparing predictive accuracy, Journal of Business and Economic Statistics 20 (2002) 134–44. `doi:10.1080/07350015.`

1995.10524599.

A. Tascikaraoglu, M. Uzunoglu, A review of combined approaches for prediction of short-term wind speed and power, Renewable and Sustainable Energy Reviews 34 (2014) 243–254. doi:https://doi.org/10.1016/j.rser.2014.03.033.
URL https://www.sciencedirect.com/science/article/pii/S1364032114001944

P. M. Bento, J. A. Pombo, S. J. Mariano, M. R. Calado, Short-term price forecasting in the iberian electricity market: Sensitivity assessment of the exogenous variables influence (2022) 1–7doi:10.1109/EEEIC/ICPSEurope54979.2022.9854716.

A. Hošovský, J. Piteľ, M. Adámek, J. Mižáková, K. Židek, Comparative study of week-ahead forecasting of daily gas consumption in buildings using regression arma/sarma and genetic-algorithm-optimized regression wavelet neural network models, Journal of Building Engineering 34 (2021) 101955. doi:https://doi.org/10.1016/j.jobe.2020.101955.
URL https://www.sciencedirect.com/science/article/pii/S2352710220335877

G. Woo, C. Liu, D. Sahoo, A. Kumar, S. Hoi, Learning deep time-index models for time series forecasting (2023). arXiv:2207.06046.