Mitigating the Weaknesses of Machine Learning in Short–Term Forecasting of Aggregated Power System Active Loads

Pekka Koponen
Power Systems and Renewables
VTT Technical Research Centre of Finland
Espoo, Finland
Pekka.Koponen@vtt.fi

Harri Niska
University of Eastern Finland
Kuopio, Finland
Harri.Niska@uef.fi

Antti Mutanen
Electrical Engineering
Tampere University
Tampere, Finland
Antti.Mutanen@tuni.fi

Abstract— **Machine learning methods predict accurately in situations that are adequately included in the learning data and do not require detailed domain knowledge based model development. They have their weaknesses compared with other forecasting methods, however. For example, they may fail in many new situations not experienced before. Hybrid models are increasingly popular as they are capable of combining the strengths of several modelling methods and mitigating the weaknesses. We study short–term forecasting of aggregated electricity demand that includes dynamically controlled thermal storage. Purely measurement data driven models tend to fail in forecasting power in rarely occurring situations, such as dynamic load control actions and extreme weather. The thermal dynamics of the loads, large outdoor temperature variations, and changes in the energy technologies contribute to this challenge. Combining various information sources and the strengths of different modelling approaches is needed. We study the following approach using field trial data covering over 7500 houses and 27 months. We forecast control responses and load saturation using models that have physically based model structures. Then we forecast the residual using data driven models, such as machine learning models designed and tuned to learn also system dynamics. The load forecast is the sum of these component forecasts. We further improve the forecast by using ensemble forecasting and physically based range forecasts. We find that the hybrid methods are more accurate than their component methods alone and combining several hybridization approaches can improve the performance and reliability.**

Keywords—forecasting, hybrid intelligent systems, machine learning, multilayer perceptrons, power demand, support vector machines

I. Introduction

Accurate and reliable forecasts of the electricity market loads, balances, and the distribution system power loading are a critical enabler for high penetrations of distributed power generation and demand response. Ignoring the explicit presence of active demand in the load model leads to unsatisfactory forecasts according to [1] and [2]. Machine learning based forecasting methods are widely applied in forecasting load and generation in power systems. Awareness of their fundamental strengths and limitations is accumulating. The limitations mainly stem from the lack of transparency and from ignoring domain knowledge and the information available in physical models.

It is increasingly popular to improve forecasting accuracy by combining different forecasting methods to hybrid methods [3]. There are many approaches for the combination. In an ensemble approach, several forecasting algorithms run in parallel and use a weighted average of the forecasts, while adjusting the weights according to the situation as learned in the identification [4]. Another alternative is to use machine learning to tune the parameters of another model, such as a dynamic nonlinear state space model. Third possibility is to apply different methods to separately measured (or estimated) load components [5]. Fourth is the sequential residuals approach where machine learning forecasts the residual of another model. ARIMA models and machine learning were combined in this way by [6]. Fifth is to limit the inputs and outputs of the machine learning models using constraints forecasted by another model. Sixth is to generate additional learning data by simulations. The above list is not exhaustive.

In the present paper, we show results on how adding methods and hybridization approaches improves the forecasting accuracy. In our case, the outdoor temperature has large variations, the load behavior nonstationary and the number of load control tests small, and we found the ARIMA models to be very inaccurate. Thus, we here use models with physically based structures to forecast the control responses, and machine learning models and a similar day method to forecast the resulting residual. The load forecast is the sum of these two component forecasts. Depending on the forecasting case, we also combine such forecasts to a simple ensemble or use a physically based range forecast to limit the final forecast values to be feasible.

The component models in our hybrid forecasting methods include 1) separate partly physically based models for three different types of active demand, 2) a similar day forecaster (SD) and 3) two relatively old machine learning methods that are a support vector machine (SVM) and a multilayer

This research was mainly done in the project Response funded by the New Energy Programme 2015-2018 of the Academy of Finland.

978-1-7281-2927-3/19/$31.00 ©2019 IEEE

perceptron (MLP). We use a genetic algorithm with sensitivity functions to optimize the structure of the machine learning models. New projects, such as VTT's internal project SAISEI, study hybrids with the state of the art methods based on deep learning, such as long short–term memory (LSTM) and convolutional neural networks (CNNs), with longer data sets, but their results will be published later. Despite their great methodological potential, and performance and robustness improvements, also the novel deep learning techniques benefit from integration with other methods when the learning data lacks adequate information on critical exceptional situations.

Based on a performance comparison [5] of data driven models, SVM and MLP seem to be good machine learning methods for our forecasting purposes. According to the literature, such as [7], SVM has many benefits, such as good accuracy and insensitivity to outliers. However, SVM is also known to be computationally inefficient [5].

Using such hybrids, we forecast the hourly interval powers for spot price based control of aggregated loads of full storage heating houses [2] and [8], and forecast the power of distribution grid area with 3–minute time resolution [9]. In the present paper, we also discuss mitigating the typical weaknesses of machine learning by combining several forecasting methods and hybridization approaches.

This contribution explains some results of the project Response [10] that studied the following research hypotheses:

1) Hybrid models combine the benefits of different load modelling approaches, thus (a) forecasting relatively accurately in different situations including also those that have not been experienced before, (b) adapt to expected and unexpected changes in load behavior, and (c) are reasonably easy and fast to maintain and update.

2) Models that combine different relevant available information sources forecast dynamically controlled aggregated load more accurately than black box models or models that are purely physically based.

II. CHALLENGES WITH THE APPLIED MACHINE LEARNING METHODS

Much new energy demand related data and models have become available from various sources such as smart metering, distribution grid automation, building automation, and new public and private databases. We experienced the following challenges in transforming the data to accurate forecasting models. 1) As new energy technologies are introduced the system behavior changes and especially the machine learning models applied in this study tend to need so much learning data that the learned model is outdated. 2) The purely data driven models often failed outside the situations included in the learning data. Typically such rare situations are critical and good forecasting accuracy is especially important during them. 3) Machine learning models lack transparency so it can be extremely difficult to anticipate how they perform in new exceptional situations. 4) The machine learning forecasts tend to be relatively sensitive to errors and outliers in the input data. This can make them vulnerable to cyber–attacks and ICT errors. 5) Crucial information was lost in the pre-processing of the identification data and not detected before the forecasting failed. 6) Existing load forecasting models typically model demand as passive and fail when forecasting in the presence of substantial amounts of active demand. 7) Modelling the time dynamics used to be rather exploratory in machine learning and the experience on different approaches is still rather limited. We found out that hybrid models mitigated these weaknesses with the studied data driven methods. The progress in machine learning may create also other potential solutions than the hybrids.

Machine learning has also relative strengths that make it useful and increasingly popular. Relatively simple physically based models are superior in forecasting active demand responses and many new or changing situations, but in forecasting the total load they require very much domain expertise and model development work and still most of the time have inferior accuracy compared with machine learning. Combinations of the different methods to hybrids helped to mitigate the above mentioned challenges.

The following challenges were similar in the machine leaning methods and in their hybrids. 8) MLP often failed to converge properly in identification thus requiring repeated runs with different parameters. 9) SVM scaled poorly to large problems. 10) It was necessary to split the identification data to separate learning data and data for controlling overfitting. In the hybrid the tendency to overfitting was clearly mitigated.

III. THE SHORT–TERM FORECASTING PROBLEMS

Short–term load forecasting typically means forecasting powers for less than 1 week ahead but at least a day ahead. Here we study forecasting at 9 a.m. the power in each time interval of the next day (15–39 hours ahead). The methods produce also forecasts for the same day and for the day after tomorrow, but the performance indicators are easier to understand, if the forecasts do not overlap each other.

The problem studied is to forecast the next day aggregated power of loads that include active demand (AD). It comprises forecasting 1) the power at 3–minute intervals for a power distribution area comprising several primary substations and 2) the aggregated power of each separately controlled group of active demand customers. The powers of the groups that include control responses are needed mainly for the electricity markets that now operate with hourly intervals and in the future at 15–minute time resolution. We forecast powers of the active demand groups with 3–minute time resolution or better which enables us to use the same response models of the active demand groups also in forecasting the power of the distribution area.

There are mainly two types of active demand in the studied distribution area. One is a Time–of–Use (ToU) service. It is now operated statically based on the time. The ToU technology in the area is capable to dynamic direct load control operation based on, for example, electricity market price variations. The other active demand type is emergency load control that provides peak load reserve for the transmission system and for the distribution grid. It is important to forecasts powers so that the responses of both types of active demand are accurately included.

In this case the AD comprises over 7500 hourly interval metered small electrically heated houses in 5 separately controlled groups. The distribution network operator can send a signal that temporarily switches off daytime heating loads and possibly also cooling loads. All the sites have also ToU controlled nigh time electrical heating.

The power of the power distribution area is measured at the primary substations with 3–minute time intervals. Hourly interval consumption measurements from the previous day are available from each customer. We used the outdoor temperature and its forecast for Kajaani, the central city of the power distribution area. The impacts from solar radiation, snow cover, wind speed and humidity are much smaller and are ignored here for clarity. In forecasting the total power of the distribution grid area, the identification period was 12 months long in 2011 and 2012, see Fig. 1, and the verification period was the two first months of the year 2014, see Fig. 2.

In forecasting the group powers, the identification period was13 months in 2011 and 2012. Then the verification period comprised 14 months in 2013 and 2014. Both the identification and verification periods included some emergency load control tests and daily time of use (ToU) control actions.

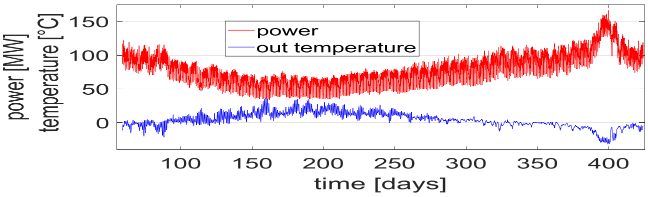

Fig. 1. Power in the distribution area in the identification period, time starts 1.1.2011 00:00.

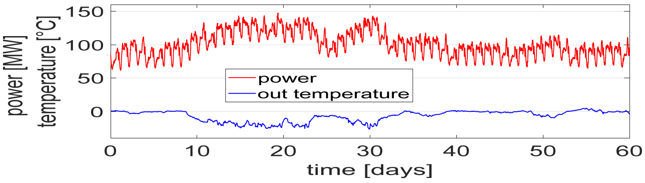

Fig. 2. Power in the distribution area in the verification period, time starts 1.1.2014 00:00.

IV. The Methods

A. The Hybrid Approaches

Fig. 3 shows the main structure of the hybrid forecasting model. The input variables include time t, outdoor temperature T_{out}, and, for every controlled group i, the AD control signal u_i, past hourly interval power P_i and the number of sites n_i. Partly physical models forecasts the AD responses for each controlled group and machine learning is taught to forecast the residual. The result is the forecast grid area power P_{f}.

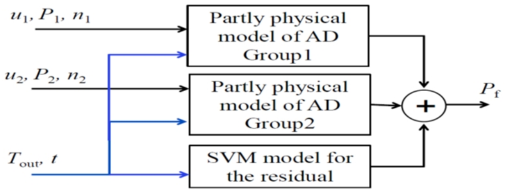
u_1, P_1, n_1
u_2, P_2, n_2
T_{out}, t
Partly physical model of AD Group1
Partly physical model of AD Group2
SVM model for the residual
+
P_{f}

Fig. 3. Machine learning forecasts the residual of the partly physically based response model.

In the Fig. 3 the residual is forecast by SVM but we forecast similarly the residual with some other data driven models that are MLP and SD. In order to improve the forecasting performance and reliability we add some other hybridization methods, see Fig. 4. There an ensemble forecasting makes a weighted average of component forecasts. Additional partly physically based models forecast daily energy demand of each controlled group and the feasible ranges of the power. The AD forecasting needs them as inputs. We also use the range forecast to limit the final forecast P_{f} to feasible values, thus improving the forecast slightly during exceptional weather situations, for example.

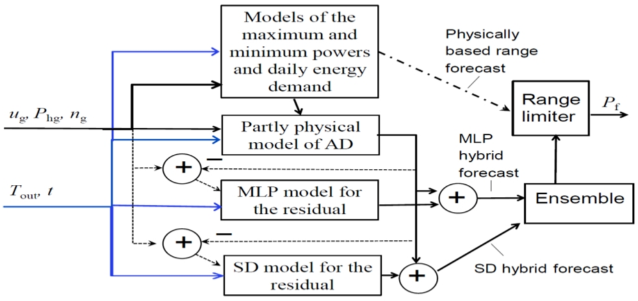
Models of the maximum and minimum powers and daily energy demand
Physically based range forecast
Range limiter
P_{f}
u_g, P_{hg}, n_g
Partly physical model of AD
MLP hybrid forecast
T_{out}, t
MLP model for the residual
Ensemble
SD model for the residual
SD hybrid forecast
+
−
+
−
+
+

Fig. 4. A good combination of three hybridization approaches that are 1) sequential residuals, 2) ensemble and 3) range forecasting.

Fig. 4 shows a block “Partly physical model of AD”. This block comprises several partly physical models in parallel, because each controlled group has its own AD control signals and consequently its own partly physical AD models. The model of each AD control group can further comprise separate models according for different types of the AD service.

B. Physically Based Emergency Control Model

The model for the emergency control responses was developed in [11] from 3–minute interval power measurements from primary distribution substations. There, we identified the parameters using non–linear constrained optimization. Constraints for parameter ranges were estimated from the building requirements of the climate zone.

Outdoor temperature T_{out}, and the group control signals u_i are the input variables. The state variables comprise four internal temperatures and one temperature controller state. Each internal temperature is associated with a thermal storage capacity.

In real houses, the temperature controls are often on–off type. The heating is either on full power or zero power. Such a model is very inaccurate in forecasting the aggregated behavior unless a large number of models with stochastic disturbances is run in parallel. Thus we use a continuous controller in the house model and after the parameter identification it gives a dynamic behavior similar to the aggregated group.

The model output power is scaled by applying a slow first order filtered feedback from P_{ih}, the measured hourly interval consumption of the group, and n_i, the number of houses in the group i. The scaling function was separately identified from the new identification data (2011 and 2012). The function needs an average site size estimate as input and gets it using slow feedback from the measurements (time constant several weeks) and an even slower forgetting factor. The resulting model accurately forecasts the aggregated responses also when the heating in the individual houses is controlled on–off.

The emergency load control tests in the identification period did not include so cold temperatures that the identification of heating power saturation would have been possible. In three neighbor power distribution networks, similar aggregated house models had been similarly identified in 1996–1997 from the primary substation power measurements. Then the tests included cold enough temperatures, 13 substations and 6 separately controlled load groups [11]. The measurements covered one year. The model of a large group had very good agreement with the responses identified in the new identification data. Thus, we here use that very old aggregated house response model as such as the partly physical model of AD in all the analysis and results related to the forecasting of the power of the distribution area.

C. Time Of Use Control Model

For modelling the Time of Use (ToU) control responses, we first identified how the energy consumed during the day depends on the outdoor temperature and how the temperature dependence differs according to the state of the ToU control signal. Also the temperature dependence of the load increase and decrease at the change of the control signal state were identified. Then we developed a simple heat storage model with first order dynamics and state constraints. It can only fit to the first hours of the control response, when the heating turns on, because in the identification data the load typically increases at the end of heating periods. Adding a load component that exponentially increases with time improves the fit during the last few hours of the heating periods. The model parameters were identified from the identification data. We prefer using minimalistic models, because forecasting the step changes is the main purpose of this ToU response model and all the data driven models (SVM, MLP and SD) applied to the residual model the other aspects of the response well enough.

D. The Machine Learning Methods

Two machine learning methods are compared: 1) support vector machine (SVM) and 2) multilayer perceptron (MLP). We used direct prediction scheme for both these machine learning models by using delayed power and temperature values as regressors. There are many alternative, often better, approaches for modelling the time dynamics of the nonlinear system. Many of them use feedback from delayed output values or internal states as inputs to the neural network. They include Kalman filter, Elman dynamic recursive network [12], hybrid nonlinear autoregressive MLP [13], long short–term memory (LSTM) [14], [15], LSTM based sequence to sequence architecture [16], combination of convolutional neural networks (CNN) with LSTM and deep neural network [17], deep residual networks [18] and gated recurrent unit neural network (GRU) [19]. Now we do not compare them, because the focus is on the benefits of integrating different models in order to include also such available crucial information that is missing from the learning data. We can expect the benefits to be generic, because methods can only extract such information that exists in its input data.

SVM is a machine learning technique for data classification and nonlinear regression. The main technical details of SVM are explained in [20]. Epsilon(ε)–SVM with the radial basis kernel function based on the LIBSVM package was used to execute the model runs.

The MLP was trained using Levenberg–Marquart modification of the back propagation. The algorithm iteratively adjusts the weights of the squared errors between the forecast and measurement outputs. One hidden layer was used. More detailed explanation is in [21].

Table I describes input variables for the machine learning models. A multi–objective genetic algorithm was applied with sensitivity analysis to select an optimal subset of the inputs variables [21]. A tedious and poorly reproducible trial–and–error effort was thus avoided. We transformed discontinuous timing variables into continuous form using trigonometric transformations. We smoothed the hour of the day to minute level indices using sliding average with one hour window.

It is necessary to control the risk for overfitting when applying machine learning models. In case of the MLP network, we adopted the standard method called early stopping (with 5% sampling of identification data). Contrary to the MLP network, SVM contains the control parameters (ε, C), which define the margin within which the error is neglected (noise) and the smoothness of the approximation, respectively. Values C=100 and ε=0.01 were used based on experimenting.

TABLE I. MACHINE LEARNING MODEL INPUTS AND THEIR PHYSICAL INTERPRETATION

Inputs to be optimally selected	**Physical sub load**
Day of the year (1– 365)	Domestic appliance seasonal rhythm
Day of the week (1– 7)	Domestic appliance weekly rhythm
Hour of the day (1– 24)	Domestic appliance daily rhythm
Day length (hours)	Lighting, radiation affected thermal load
Outdoor temperature (ºC) with time–lags of 1– 48 hours	Thermal load (heating and cooling)

E. The Similar Day Forecaster

We developed a similar day forecaster for the load in the studied distribution area. In this forecaster, the load on each 3–minute interval is forecast based on earlier intervals with similar characteristics. Typically, there are several similar intervals and average load on these intervals is used. Table II gives the considered characteristics and their averaging windows.

Unlike in some similar day forecasters presented in literature, such as [22] and [23], we modelled the load's dependency on the outdoor temperature separately. The temperature dependency (W/ºC) for each forecasted day is determined with simple linear regression. The effects of intra–week fluctuations in electricity demand are eliminated by choosing the dependent and independent variables as follows: Dependent variable is the difference between the daily energy consumption and the average daily energy consumption on similar days of the week. Independent variable is the difference between the daily average of effective hourly temperatures and the average of effective hourly temperatures on similar days of the week. The effective hourly temperature is defined as an average over the previous 24 hourly temperatures. In regression analysis, the effects of seasonal fluctuations are eliminated by using data from only similar days of the year. The identified temperature dependency is then used to correct the average load of similar intervals to correspond the load in the forecasted temperature. Finally, the systematic forecasting errors, possibly caused by rising or falling trends in electricity consumption, are corrected based on (uncorrected) forecasting errors on the preceding 30 days.

TABLE II. CHARACTERISTICS DEFINING THE SIMILAR INTERVAL

Characteristic (index range)	**Size of the moving averaging window**
Day of the year (1– 365)	±15 days from the identification data and previous 15 days from the verification data
Day of the week (1– 7)	0 (Must be exactly the same)
Public holiday or other special day (0– 17, 0=normal day)	0 (If index>0, day of the year and day of the week are ignored)
3–minute interval of the day (1– 480)	±1 interval
ToU control signals (0 or 1, two control signals)	0

V. PERFORMANCE INDICATORS

We use Root Mean Square Error (RMSE) in MW as a performance criterion for forecasting accuracy both in identification and in performance comparisons. In order to enable comparison between different forecasting tasks we also give the Normalized Root Mean Square Error (NRSME), which is the RMSE normalized to the mean power of the identification or verification period respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t)^2} \qquad (1)$$

$$NRMSE = RMSE/\bar{y} \quad , \qquad (2)$$

where N is the total number of data points, y_t is the measured value at time point t, $\hat{y}_t$ is the forecast value at time point t, and $\bar{y}$ is the mean of the measured data.

In this forecasting task, accurate forecasts during peak power are important, because then large forecasting errors lead to especially high costs in energy markets and in grid operation. Then large forecasting errors are especially expensive. The criterion must respectively penalize outliers in the forecasting errors during high loads. Thus, for example, the mean absolute percentage error (MAPE) is not a suitable forecasting performance criterion for our load forecasting case and we do not use it any more. Even NRMSE may not adequately weight the forecasting errors during high loads and additional criteria such as NMRSE for the 5 % or 10 % of the highest loads could be useful.

Forecasting accuracy is not alone an adequate measure for the comparison of forecasting methods. Reliability, and data and computation cost are suggested in [24], which also points out that scale independence is needed to enable comparison across models and applications. Developing and comparing forecasting methods needs that even more aspects are considered. An ideal method

- is easy to understand, develop and maintain and
- robust to errors in input data and to cyber–attacks,
- takes care that the forecasts are feasible,
- maintains reasonable accuracy also in rare and new situations not included in the learning data
- remains accurate also in presence of dynamic AD,
- does not require excessively long history for learning, and
- adapts to changes in the load behavior,

Hybrid methods provide solutions to these challenges, but the comparison is difficult without quantitative common metrics.

VI. RESULTS

A. Active Demand

The partly physically based AD forecasts at an emergency control action in the verification are in Fig. 5. They represent the aggregated load control responses of over 7500 houses. The sum of the ToU forecasts is blue and the sum of the emergency control forecasts is red. Summing all of them up, gives the total AD forecast. The full verification period and the total power to be forecast are shown in the Fig. 2. The residual is large and in the hybrids the data driven models are forecasting it. The hybrids forecast accurately the load during the control actions but the machine learning methods, MLP and SVM, and SD fail completely. Fig. 6 shows this for the MLP. Fig. 7 shows the forecasts of the different hybrid models that comprise the partly physical response models and the residual forecasting data driven models studied.

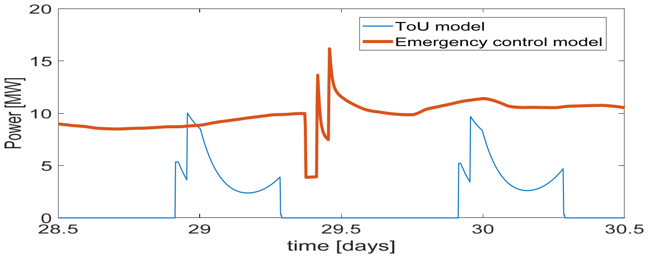

Fig. 5. A sample of the partly physically based response forecasts of the power distribution area in verification.

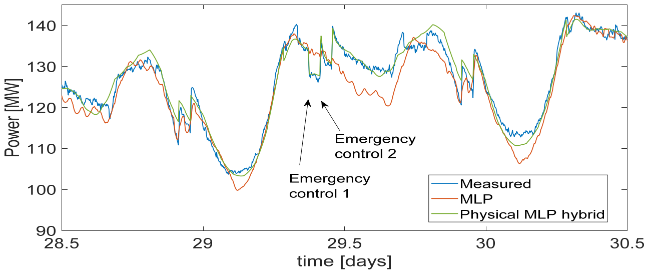

Fig. 6. Physical MLP hybrid forecasts the load control responses but MLP fails.

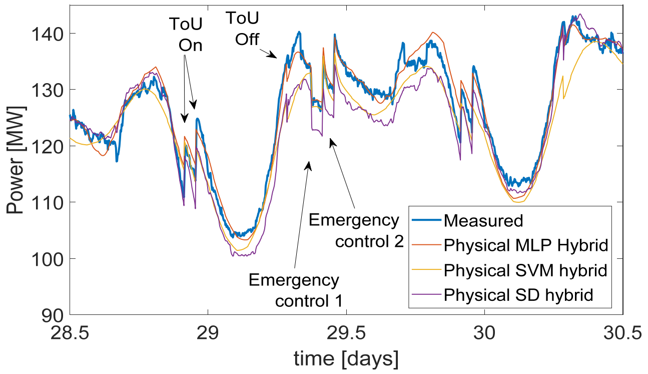

Fig. 7. A sample of the verification of hybrids that combine partly physical models to data driven models in forecasting the power of the distribution area.

The ToU loads were turned on in two steps. The emergency control actions were one hour long and started at two different times separated by one hour. The hybrid models where the data driven models forecast the residual of the partly physically based control response models forecast the AD control responses accurately.

B. *Combining Forecasting Methods Improved Accuracy*

Table III gives an overview of the forecasting performance of the compared methods and method combinations in forecasting 3–minute interval power of the electricity distribution network area in the Figs. 1 and 2 using the approach in the Fig. 3. In the Table III, combining many different forecasting models and hybridization methods improves the accuracy of the forecasts substantially. The performance of the machine learning methods and hybrids depends on the parameters that tune the learning process. The most accurate methods are hybrids that have AD models for each AD type and group, residual forecasting by both SD forecaster and either one of the machine learning methods combined to a very simple ensemble that calculates the mean of the residual approach based hybrids.

The SD forecaster is the only one of the methods that includes a model for special days. That explains to some extent why it is more accurate than the machine learning methods (MLP and SVM) and also included in all the best hybrid methods. Another explanation is that the MLP and especially the SVM smooth out the regular rapid power changes that result from the ToU load control.

The results with the MLP depend on the starting weights used in the identification. The SVM results depend on the random sampling. Thus their results vary slightly from run to run. The values in the Table III are typical rather than the best results.

In the last rows of the Table III using simple ensemble improves the performance further provided that the same model (here SD) is not applied repeatedly on the different phases of the hybrid, compare the rows simple ensemble 1.1 with 1.2, and 2.1 with 2.2. However, parallel use of the control response models in the ensemble components gives the best results. The performance of the final ensemble in verification is here better when the less accurate MLP and SVM hybrids are used. In the identification, the ensemble forecasts are, of course, more accurate when more accurate component forecasts are applied. Tuning the ensemble to best identification performance puts very much weight to the better one of its component hybrid forecasts, but in the verification the best performance is achieved by putting almost equal weight to both component hybrid forecasts. That can be expected, because different methods complement each other.

TABLE III. COMPARISON OF METHODS IN FORECASTING THE POWER OF THE DISTRIBUTION AREA WITH 3-MIN TIME RESOLUTION

RMSE in MW and (NRMSE in %)	**Identification**	**Verification**
MLP	2.5936 (3.34)	4.7783 (4.60)
SVM (5% sampling)	3.4444 (4.43)	5.2658 (5.07)
Similar day forecaster (SD)	2.3158 (2.98)	3.7466 (3.71)
Emergency control response model and MLP residual model	2.5603 (3.30)	4.2772 (4.11)
Emergency control response model and SVM residual model (5%)	3.4159 (4.40)	5.2414 (5.04)
Emergency control response model and SD residual model	2.2556 (2.90)	3.6427 (3.61)
Emergency control response and ToU response models and residual MLP (MLP hybrid)	2.6869 (3.46)	3.8647 (3.81)
Emergency control response and ToU response models and residual MLP (MLP hybrid)	2.8327 (3.65)	4.3488 (4.28)
Emergency control response and ToU response models and residual SVM (5% sampling) (SVM hybrid)	2.2355 (2.87)	3.5342 (3.50)
SD, emergency control response and ToU response models (SD hybrid)	2.5936 (3.34)	4.7783 (4.60)
SD, emergency control response and ToU response models + MLP (SD MLP RM hybrid)	1.7299 (2.23)	3.5525 (3.50)
SD, emergency control response and ToU response models + SVM (SD SVM RM hybrid)	1.6724 (2.15)	3.5014 (3.45)
Simple ensemble 1.1 (weighted mean of the MLP hybrid and the SD hybrid)	1.9838 (2.55)	3.0913 (3.05)
Simple ensemble 1.2 (weighted mean of the SD MLP RM hybrid and the SD hybrid)	1.7299 (2.23)	3.4189 (3.29)
Simple ensemble 2.1 (weighted mean of the SVM hybrid and the SD hybrid)	2.0034 (2.58)	3.0140 (2.97)
Simple ensemble 2.2 (weighted mean of the SD SVM RM hybrid and the SD hybrid)	1.6724 (2.15)	3.5014 (3.45)
Very simple ensemble 2.1 (mean of the SVM hybrid and the SD hybrid)	2.0676 (2.66)	3.0681 (3.02)

C. *Range Forecasting To Keep The Outputs of Data Driven Forecasts Feasible*

The machine learning methods studied tend to produce forecasts that are not completely feasible during AD events, extreme weather conditions and other situations that are rare or missing in the identification data. Thus, we often forecast the feasible output range separately and use it to limit the outputs of the machine learning methods. We use both partly physically methods and data driven methods for the purpose.

We explain in [8] a case, where we forecast aggregated powers of groups of full storage houses. The total number of houses is over 700. Large hot water tanks are used as the heat storage controlled according to the electricity market spot price. There, we use hybrids comprising a partly physical control response model and a machine learning method. The problem is that the machine learning both generalizes and extrapolates the load behavior so that even the hybrid forecasts include slightly infeasible values. The forecasts include negative powers and power peaks that exceed the possible or likely maximum aggregated load for the weather conditions. Thus, we separately forecast the daily minimum and maximum loads using either data driven or simple physically based methods depending on which ones turn out to be the most accurate. The improvement in the RMSE is rather small, but the range limiting makes the forecast more reliable and robust in situations that are new, rare or include large errors in the input data of the forecasting model. Table IV shows the results. The range limiting model includes a physically based nonlinearity.

TABLE IV. LIMITIG THE FORECAST TO FEASIBLE RANGE

RMSE kW/house (NRMSE %)	**Identification**	**Verification**
partly physical without response model	3.171 (98.83)	2.646 (113.71)
partly physical with response model	1.0753 (33.51)	1.219 (52.39)
response model and SVM	0.7326 (22.72)	0.8486 (36.04)
response model, SVM and minimum	0.7308 (22.78)	0.8042 (34.27)
response model, SVM and range limit	0.7304 (22.76)	0.8031 (34.24)
SVM	0.5509 (17.17)	1.744 (74.93)

In this full storage electrical heating case, we got the best accuracy by using in the hybrids a machine learning model (SVM) and two different partly physically based models. One forecasts the heating load dynamics and control responses, and the other forecasts the feasible range of the main forecast. The NRMSE values are high, because the loads consume most of their energy in load peaks that are high compared to the mean power taken by the houses. This demonstrates the fact that even when using so called scalable metrics, NRMSE and MAPE, comparison across applications is not straightforward.

The SVM clearly overlearned due to the short identification period but the hybrids with the same SVM structure did not. This suggests that the identification period was too short for the SVM. Increasing the length of the identification period can be included in future studies, because EU H2020 project SysFlex amended the test data set by 3.5 new years and now studies the hybrid model concept in forecasting also the amount of controllable power.

For one week of the verification period, the measured load and the forecast by the hybrid including a response model, SVM and the minimum limiter of the Table IV are shown in Fig. 7. The high load peaks are due to the control signal that turned on the storage to meet the forecast daily energy demand with minimum costs when subject to day ahead spot market hourly prices and a ToU grid tariff. The minimum limit forecast has limited some negative values to small positive values. Tuning the minimum limit forecast slightly higher would have slightly improve the forecast accuracy further. In the forecast there are two so high peaks that applying the upper limits of the range limiters would have cut them, if applied. The load in Fig. 7 comprises two separately controlled load groups. There were not any large outliers in the residual as can be seen from Fig. 8.

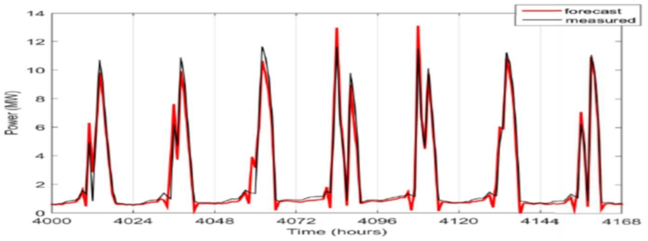

Fig. 7. The dynamically controlled full storage heating load and its hybrid forecast in the verification [2].

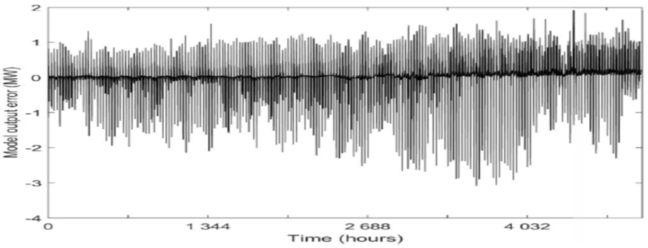

Fig. 8. The residual of the hybrid forecast during the 7–month long verification period of the full storage heating case [2].

In Table IV the range limit includes a forecast maximum limit in addition to the forecast minimum limit. The range limit takes care that the forecasts always remain in reasonable range in any situation and thus mitigates a significant machine learning vulnerability.

In Table IV and Figs. 7 and 8 the RMSE was used as the model identification criterion. Using MAPE as the identification criterion resulted in even higher exaggeration of the load peaks in the forecasts. Omitting the partly physically based control response model also increases the overestimation of the height of the highest load peaks. This evidence suggests that a separate model for forecasting the load range and limiting the load forecasts is justified.

VII. DISCUSSION

Our results consistently show evidence supporting the combination of several load forecasting methods by using many hybridization approaches. Further research is needed to better understand how and to what extent hybrid models can mitigate the weaknesses of machine learning models, such as:

- development of test cases and additional complementing performance criteria,
- analysis of more forecasting cases in order to understand to what extent the findings are generic,

- using the experience from the SD method in selecting the delayed inputs from measured power to the ML methods.
- including new machine learning and deep learning methods for learning system time dynamics
- adding online learning.

The hourly interval smart metering data is subject to data privacy legislation and cannot be made publicly available. The access to the prepossessed and aggregated data and grid data is to be decided together by the relevant distribution network operators and the research organizations.

VIII. Conclusions

Integrating many forecasting models and hybridization approaches improves forecasting accuracy in the studied short–term load forecasting cases. Each methods has its strengths and weaknesses. The hybrid models combine the strengths and mitigate the weaknesses. It is increasingly important that load forecasts remain accurate also when demand is active and subject to dynamic control actions that are based on the situation rather than the clock. Data driven models, including machine learning, are poor in forecasting such responses of dynamic load control and other rarely occurring situations. In such situations, they may even produce forecasts that include infeasible values. Hybrid models offer solutions to these challenges.

Forecast accuracy is only one of many requirements for the selection of forecast models. The hybrids also offer advantages regarding other requirements for the models. These additional advantages are difficult to quantify and compare, because common quantitative metrics suitable for the particular forecasting task still need to be developed.

Acknowledgment

The electric power companies Helen and Loiste implemented the field tests and provided the measurements and information for the modelling. This research would not have been possible without that. Data and models from earlier field tests by five other electric power companies were also used. We received weather data both from the electric power companies and the Finnish Meteorological Institute.

References

[1] A. Garulli, S. Paoletti, A. Vicino, ”Models and Techniques for Electric Load Forecasting in the Presence of Demand Response”, *IEEE Transactions on Control Systems Technology*, Vol. 23, No. 3, May 2015, pp. 1087–1097.

[2] P. Koponen, H. Niska, “Hybrid Model for Short–Term Forecasting of Loads and Load Control responses”, *IEEE PES ISGT Europe 2016*, 6 p.

[3] C. Deb, F. Zhang, J. Yang, S. E. Lee, K. W. Shah, “A review on time series forecasting techniques for building energy consumption”, *Renewable and Sustainable Energy Reviews* 74 (2017), pp. 902–924.

[4] B. Cs. Csáji, A. Kovács, J. Váncza,”Online Learning for Aggregating Forecasts in Renewable Energy Systems”, *ERCIM NEWS 107*, October 2016, pp. 40–41.

[5] B. Dong, Z. Li, S. M. M. Rahman, R. Vega, “A hybrid model approach for forecasting future residential electricity consumption”, *Energy and Buildings* 117 (2016) 341–351.

[6] O. Valenzuela, I. Rojas, F. Rojas, H. Pomares, L.J. Herrera, A. Guillen, L. Marquez and M. Pasadas, “Hybridization of intelligent techniques and ARIMA models for time series prediction”, *Fuzzy Sets and Systems*, vol. 159, Elsevier 2008, pp. 821–845.

[7] A. Selakov, S. Ilic, S. Vukmirovic, F. Kulic and A. Erdeljan, “A Comparative Analysis of SVM and ANN Based Hybrid Model for Short Term Load Forecasting”, *Transmission and Distribution Conference and Exposition*, *2012 IEEE PES*, 7–10 May 2012, Orlando, FL. 5 p

[8] P. Koponen, H. Niska, R. Huusko, “Improving the performance of machine learning models by integrating partly physical control response models in short–term forecasting of aggregated power system loads”. *ITISE 2017*, 12p

[9] P. Koponen, H. Niska, A Mutanen, “Combining the strengths of different load modeling methods in short–term load forecasting of a distribution grid area with active demand”, *CIGRE D2 Colloquium* 2019, 8p.

[10] P. Koponen, S. Hänninen, A. Mutanen, A. Rautiainen, J. Koskela, H. Koivisto, Pertti Järventausta, H. Niska, M. Kolehmainen, “Improved Modelling of Electric Loads for Enabling Demand Response by Applying Physical and Data–Driven Models, project RESPONSE”, *IEEE ENERGYCON* 2018. 6 p.

[11] P. Koponen, *Optimisation of load control*, Final report, VTT Energy. Espoo, 20 November 1997, 26 p. + app. 14 p. Research report ENE6/12/97, 26 p. + app. 14 p.

[12] Feng Zhao, Hongsheng Su, “Short-Term Load Forecasting Using Kalman filter end Elman Neural network”, *2nd IEEE Conference on Industrial Electronics and Applications*, 2007, pp. 1043–1047.

[13] Gang Lv, Zhiming Liu, Yu Fan, Guo–Guo Li, “Modeling a permanent–magnet linear synchronous motor using hybrid nonlinear autoregressive neural network”, 2008 *9th International Conference on Signal Processing*, 2008, pp. 1685–1689.

[14] S. Hochreiter, J. Schmidhuber, “Long short–term memory”, *Neural Comput*. 1997, 9, 1735–1780.

[15] Seon Hyeog Kim, Gyul Lee, Gu–Young Kwon, Do–In Kim, Yong–June Shin, “Deep Learning Based on Multi–Decomposition for Short–Term Load Forecasting, *Energies* 2018, 11, 3433, pp. 65–81.

[16] D. L. Marino. K. Amarasighe, M. Manic,”Building Energy Load Forecasting Using Deep Neural Networks”, *IEEE IECON 2016*, pp 7046-7051.

[17] Tae-Yong Kim, Sung-Bae Cho, “Predicting the Household Power Consumption Using CNN-LSTM Hybrid Networks”, *IDEAL 2018* Proceedings, Springer, pp. 481-490.

[18] Kunjin Chen, Kunlong Chen, Qin Wamg, Ziyu He, Jun Hu, Jingliang He, “Short-term Load Forecasting with Deep Residual Networks”, paper submitted on May 2018 to IEEE Transactions pn Smart Grid, 10 p.

[19] Yixing Wang, Meiqin Liu, Zhejing Bao, Senlin Zhang, “Short–Term Load Forecasting with Multi–Source Data Using Gated Recurrent Unit Neural Networks”, *Energies* 2018, 11, 1138, pp. 372–390.

[20] V. N. Vapnik, *The nature of statistical learning theory*. New York: Springer, 1995, 188 p.

[21] H. Niska, P. Koponen, A. Mutanen, “Evolving Smart Meter Data Driven Model for Short–Term Forecasting of Electric Loads”. *IEEE ISNNIP 2015*, Singapore, 7–9 April, 2015, 6p.

[22] Qingqing Mu, Yonggang Wu, Xiaoqiang Pan, Liangyi Huang and Xian Li, "Short–term load forecasting using improved similar days method," 2010 *Asia–Pacific Power and Energy Engineering Conference*, Chengdu, 2010, pp. 1–4.

[23] M. Karimi, H. Karimi, M. Gholami, H. Khatibzadehazad and N. Moslemi “Priority index considering temperature and date proximity for selection of similar days in knowledge–based short term load forecasting method”, *Energy*, Vol. 144, Feb. 2018, pp. 928–940.

[24] S. Aman, Y. Simmhan, V. K. Prasanna, “Holistic Measures for Evaluating Prediction Models in Smart Grids”, *IEEE Trans. on Knowledge and Data Engineering*, Vol 27. Issue 2, Feb. 1 2015, pp. 475–488.