



Daily middle-term probabilistic forecasting of power consumption in North-East England

Roberto Baviera¹ · Giuseppe Messuti¹

Received: 28 July 2021 / Accepted: 20 March 2023
© The Author(s) 2023

Abstract

Probabilistic forecasting of power consumption in a middle-term horizon (few months to a year) is a main challenge in the energy sector. It plays a key role in planning future generation plants and transmission grid. This paper proposes a novel model that (i) incorporates seasonality and autoregressive features in a traditional time-series analysis and (ii) includes weather conditions in a parsimonious machine learning approach, known as Gaussian Process. Applying to a daily power consumption dataset in North East England, provided by one of the largest energy suppliers, we obtain promising results in Out-of-Sample density forecasts up to one year, even using a small dataset, with only a two-year calibration set. For the evaluation of the achieved probabilistic forecasts, we consider the pinball loss—a metric common in the energy sector—and we assess the coverage—a procedure standard in the banking sector after the introduction of Basel II Accords—also running the *conditional* and *unconditional tests* for probability intervals. Results show that the proposed model outperforms benchmarks in terms of both accuracy and reliability.

Keywords Power consumption · Probabilistic forecast · Middle-term · Machine learning · Gaussian process

JEL Classification C14 · C51 · C53 · Q47

Notation and abbreviations

Notation

Symbol	Meaning
Y_t	Log-scaled power consumption time-series
y_t	Realized OS log-scaled power consumption time-series

✉ Roberto Baviera
roberto.baviera@polimi.it

Giuseppe Messuti
giuseppe.messuti@mail.polimi.it

¹ Politecnico di Milano, P.zza L. da Vinci 32, 20133 Milan, Italy

t	Cardinality of the observation
T_t	Trend
S_t	Seasonality
R_t	Residual at time t , t th component of the vector $(\mathbf{R}, \mathbf{R}_*)$
\mathbf{R}	Vector of IS residuals $\in \mathbb{R}^n$
\mathbf{R}_*	Vector of OS residuals $\in \mathbb{R}^{n_*}$
$\{\beta\}_{i=0,\dots,5}; \gamma$	GLM parameters
$D_{Sat}(t)$	Dummy variable for Saturdays
$D_{Sun}(t)$	Dummy variable for Sundays
n	IS time-series length
n_*	OS time-series length
$\mathcal{N}(\cdot, \cdot)$	Gaussian multivariate distribution
$\sigma, \sigma_f, \sigma_l$	Parameters of GP kernel
$R(x)$	Residual of Gaussian process
$k(x, \bar{x})$	Kernel (or correlation function) of \mathcal{GP}
δ_{ij}	Kronecker delta, which is one iff $i = j$ and zero otherwise
$H(X, X)$	Covariance matrix $\in \mathbb{R}^{n \times n}$ equal to $K(X, X) + \sigma^2 I$
$K(\cdot, \cdot)$	Covariance matrix $\mathbb{R}^{n \times n}$ built through $k(x, \bar{x})$
I	Identity matrix in $\mathbb{R}^{n \times n}$
x	Generic weather conditions' vector
$\ \cdot\ $	Standardized Euclidean distance

Abbreviations

<i>AR</i>	Auto regressive model
<i>ARX</i>	Auto Regressive eXogenous model
<i>GLM</i>	General linear model
<i>GP</i>	Gaussian process
<i>GPX</i>	Gaussian process eXogenous model: the proposed model
<i>iff</i>	If and only if
<i>IS</i>	In-sample
<i>LR</i>	Likelihood ratio
<i>ML</i>	Machine learning
<i>OLS</i>	Ordinary least squares
<i>OS</i>	Out-of-sample
<i>SED</i>	Standardized Euclidean distance
<i>w.r.t.</i>	With respect to

1 Introduction

Power consumption forecast has received significant attention from both academics and practitioners in recent years. In particular, middle-term forecast, i.e. in a time-horizon between a few months and a year,¹ plays a key role in the planning of power systems both for network reliability and for investment strategies in future generation plants and transmission (see, e.g., [10]).

Through a probabilistic forecast, one obtains the full probability distribution of future consumption. It is the latest frontier of current research: it is more helpful for utilities and grid operators than point consumption forecast. In fact, it does not provide only the expected value of the forecast but also information in terms of the dispersion of the forecast: a piece of information that is relevant for generating reliable scenarios.

The number of studies in the literature on density forecast for power consumption is rather limited. This technique gained momentum in the energy sector after the Global Energy Forecasting competition 2014 (GEFCOM) on a USA dataset, that includes a 7-year time-series with hourly power consumption and temperatures (dry and wet bulb) as explanatory variables (see, e.g. [11]). Some techniques have been developed focusing mainly on the short-term probabilistic forecasting of power consumption (see, e.g., [10], and references therein). One interesting probabilistic forecast has been considered by Liu et al. [16] with GEFCOM time-series.² The authors apply a quantile regression average (QRA) technique to create 24 h ahead forecasts on a rolling basis. The authors show that the QRA is able to perform, depending on the chosen metrics, 10–20% better forecasts than comparable benchmarks based on the Tao model [12]. Other recent applications of probabilistic forecast to the energy sector can be found in [26, 30].

In particular, it is important to select the relevant drivers in the forecast and their relationship with power consumption; this allows to understand and to hedge the risks. We are interested in the impact of weather conditions because they play the most relevant role in middle-term forecasts compared to economic and demographic drivers that play a role in longer forecasts [13, 14]. We focus on a region in the UK with relatively homogeneous weather conditions and on the power consumption of one main operator on the household energy market. We aim to model the dependency of power consumption from a multi-meteorological information; the most natural technique, now standard in power consumption forecasting, is known as ex-post forecasting. It has been applied to middle-term probabilistic forecasting of power consumption on the French distribution network [9], on the National Electricity

¹ In the power consumption literature, short-term consumption forecasting is the prediction of the consumption (either of one operator or of the whole system) over an interval ranging from one day to a few weeks, while long-term consumption forecasting focuses on time-horizons longer than 1 year.

² Six years of this dataset have been studied, dividing it into a 3-year training set, a 2-year validation set and a 1-year test set for model evaluation. Nowadays, it is standard in load forecasting, due to the relevant yearly seasonality, to divide training/validation/test sets into integer multiples of 1 year.

Market of Australia [14] and on one of the largest ensembles of power distribution cooperatives in North Carolina, USA [13].

We use a Machine Learning (ML) technique. In the literature, after the seminal paper of Park et al. [25] these techniques have been shown to provide interesting results in short-term point consumption forecast (see, e.g., [7, 19, 28]) furthermore, the main successes of ML have been shown with big-data analytics (see, e.g., [18, 20]). Although forecasting by nature is a stochastic problem, many studies are still using point forecasts instead of probabilistic forecasts for load, whereas probabilistic forecasting is becoming more common for pricing (see, e.g., [23], for a review). In this paper, we apply a ML technique also to a middle-term consumption forecast up to one-year; moreover, we consider a density forecast instead of a point forecast.

Among ML techniques, Gaussian processes (GP) are a natural modeling tool for density forecasting (see, e.g., [27], for a review on GP). Mori and Ohmi [22] and Yu et al. [31] have applied GP in short-time load forecasts of power consumption. In their seminal work, Mori and Ohmi [22] applied GP for one-day ahead daily maximum load forecasting on a 4-year daily load time-series of a Japanese power company. The maximum load problem allowed them to focus on the weekdays during the summer period (June–September), neglecting seasonality effects. More recently, Yu et al. [31] studied the intervention estimation problem for demand side management, analyzing a dataset obtained from a USA corporation (the Pacific Gas and Electric Company) over a 1 year and half timespan. They focused on the forecasting performance at an individual customer level, showing a mean absolute percentage error (MAPE) always above 30% even considering short time forecasts (few weeks).

Both studies have had the merit to show that GP is an attractive technique in probabilistic load forecasting with short time-series, thanks to GP parsimony. Managing short time-series is a common exigence in the industrial sector: it is quite hard to obtain, within a middle-to-large operator, homogeneous long time-series. This fact is due both to the rapid changes that are observed in this energy market and to mergers and acquisitions, corporate transactions that have become frequent after the liberalization of the sector. These power market features change significantly the composition of clients' portfolio over time and then the characteristics of the dataset.

The real challenge is to use these techniques with short time-series for middle-term forecasting: we consider the extreme case where we analyze a 2-year In-Sample dataset to forecast a 1-year power consumption time-series. Our goal is to introduce a new hybrid linear-GP model that extends the existing literature on GP (short-term) load forecasting in the following directions.

First, both Mori and Ohmi [22] and Yu et al. [31] take into account only marginally seasonal effects in their GP model: unfortunately, yearly seasonality cannot be neglected in middle-term forecasting, because the average consumption in one season can be even three times larger than the consumption in a different one. A hybrid linear-GP model allows us, first, to deseasonalize and to remove the daily autocorrelation term with a linear model, and then, to forecast the residuals with a GP.

Second, in both studies, the authors model directly the load with a GP: thus, the forecast distribution presents even negative power consumption—with finite probability—over several days. In middle-term forecasts, we forecast consumption up to 365 days after the last load data in the training set. This forecasting technique uses

previous days' forecast after the training set; thus, negative values can propagate unrealistic loads to the whole yearly forecast. In this study, we model the logarithm of the load; thus, not only we reduce the impact of seasonality (see, e.g., [10]), but also we eliminate the possibility of such unrealistic values for power consumption.

Third, both Mori and Ohmi [22] and Yu et al. [31] focus only on temperature as explanatory variable, whereas we consider a larger set of explanatory variables based on multi-meteorological data, a piece of information particularly relevant with small time-series.

Finally, both studies evaluate performances—comparing GP with the benchmarks in the test set—with a point forecast measure for the mean of the forecast (such as MAPE).³ It is true that MAPE is still a widely used error statistics in load forecasting; however, let us point out that it is very important the evaluation of the distributional features in the probabilistic forecast, when comparing the proposed model with the benchmarks. In this study, we check that not only the proposed linear-GP model performs well with point forecast measures (e.g. with a MAPE below 5% over 1 year), but also we verify the quality of the forecast from a probabilistic perspective. In this study, we consider the most important method to check the effectiveness of the probabilistic prediction: we apply a backtesting technique, standard in the banking sector after the introduction of Basel II BCoBS [1], that verifies the probabilistic features of the realized values—in terms of frequency and clustering of the events—outside the predicted confidence interval at a given nominal level. We also measure the pinball loss, a sharpness measure that is now popular in the energy sector, after GEFCom, where it was used as the scoring function in the probabilistic forecast competition [11].

The main contributions of the paper are threefold. First, we introduce a hybrid model that joints the advantages of linear time-series analysis and simple ML techniques. In particular, via a Gaussian process we incorporate the dependency from weather conditions in power consumption density forecast and we deduce density characteristics for the hybrid model. Second, we show that a hybrid linear-GP technique relying on a small dataset can achieve promising results in middle-term forecasts of power consumption. Third, we value the density forecast via both reliability and sharpness measures, pointing out the quality of the achieved results.

The rest of the paper is organized as follows. In Sect. 2, we summarize the key characteristics of the dataset. In Sect. 3, we outline the proposed model and how the weather conditions are introduced via a Gaussian process. In Sect. 4, we present the methodology; in particular, we describe (i) the forecasting technique, (ii) the benchmark models and (iii) the evaluation methods. Section 5 shows the main numerical results and Sect. 6 concludes. In a dedicated appendix, we report the notation and abbreviations used in the paper.

³ It is interesting to observe that, after having calibrated the model in the training set, they compare model performance only in the test set, without a model selection stage in a validation set: this difference with standard ML techniques (cf. e.g. [8], and references therein) is due to the absence of true hyper-parameters within GP and it is crucial with short time-series.

Table 1 Descriptive statistics for daily power consumption and average daily weather data in NE England in the whole time window

	Min	Max	Mean	Median	Standard deviation
Consumption [MWh]	248.364	1208.705	632.921	638.327	249.996
Temperature [$^{\circ}\text{C}$]	-1.330	21.581	8.902	9.020	4.520
Wind speed [m/s]	0.667	10.219	3.400	3.000	1.624
Precipitations [mm]	0.000	1.471	0.086	0.015	0.162
Chill [$^{\circ}\text{C} \cdot \text{m/s}$]	0.440	140.509	30.995	23.823	24.515
Solar radiation [kJ/m^2]	14.583	1212.917	397.424	327.083	296.357
Humidity [%]	60.167	100.000	85.415	85.344	7.525
Cloud cover	2.071	8.000	5.741	5.771	1.221

2 Dataset description

North East England is one of the nine regions of England and the eighth most populous conurbation in the United Kingdom. The dataset we analyse contains both the time-series of daily power consumption values and seven daily average weather conditions. It is three years long, from April 2014 to March 2017.

Power consumption is the aggregated household consumption of one of the main UK power suppliers. The weather dataset represents the daily average of hourly records of weather conditions in North East England. It includes seven different weather indicators:

- Temperature, in $^{\circ}\text{C}$;
- Wind speed, in m/s ;
- Precipitation amount, in mm ;
- Chill,⁴ in $^{\circ}\text{Cm/s}$;
- Solar radiation, in KJ/m^2 ;
- Relative humidity, in %;
- Cloud cover, on a scale 0 (clear) to 8 (completely cloudy).

Table 1 contains descriptive statistics about daily power consumption and weather data for the whole time window.

We notice a wide variety of units of measurement in weather conditions: it implies ranges of values that can differ by orders of magnitude. Due to the difference in units of measurement, in order to obtain a homogeneous dataset it is often useful to standardize each input variable. In the GP described in the next section, we use the standardized Euclidean distance: a simple tool that allows to consider the multi-meteorological information in the dataset.

⁴ Chill is obtained combining temperature and wind speed; it represents how cooler one feels depending on the strength of wind (see, e.g. [24]).

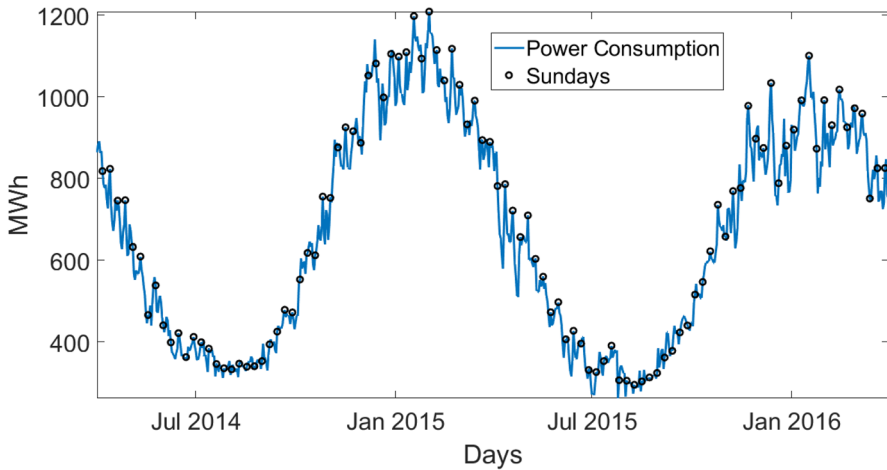


Fig. 1 Power consumption (blue line) since April 2014 to March 2016. One can notice a yearly seasonal behavior and locally higher values on Sundays (marked with a black circle) (color figure online)

The yearly and weekly seasonality of power consumption is rather evident in the dataset. Figure 1 represents the first two-year daily power consumption. It suggests a yearly seasonal behavior and slightly higher power consumption during weekends.

We observe that (i) power consumption is more than three times larger in winter w.r.t. summer time, (ii) consumption on weekdays is lower than on Sundays (and to a lesser extent on Saturdays) and (iii) volatility is larger in winter than in summer time. These are well known *stylised facts* common to most power consumption time-series: for these reasons, and in particular due to the observed volatility behavior, it is now standard to model the logarithm of power consumption (see, e.g., [10], for a review). As already emphasized in the introduction, modeling the logarithm eliminates the possibility of unrealistic negative consumption values for the forecast. In the next Section, we describe in detail the hybrid linear-GP model and in Sect. 4 the adopted methodology.

3 The model

The main goal of this study is forecasting future power consumption over the middle-term horizon, modeling both seasonal and weather-related features. Our model choice realizes a density forecast with a parsimonious description without the need to resort to extremely fine-tuned models.

We model log-scaled daily power consumption data. The characteristics of power demand that we model are:

1. long-term trend;
2. yearly and weekly seasonal behavior;
3. daily autocorrelation;

4. the relation with weather conditions;
5. weather-based error correlation and variance clustering.

In the energy literature, it is quite common a two step hybrid model approach, which takes into account, first, trend, seasonality and AutoRegressive (AR) components, and then analyses separately the residuals (see, e.g., [2], and references therein). Our hybrid model is split into two parts: a linear model (cf. Sect.3.1) for the first three characteristics (trend, seasonality and autocorrelation) and a GP (cf. Sect.3.2) that takes care of the weather influence. We call this hybrid model GPX, because it is the natural extension of AR eXogenous models, known as ARX (see, e.g. [3, p. 534 et seq.]).

3.1 The linear model part

The relation between consumption and calendar variables is established through a general linear model (GLM). The model of the natural logarithm of power consumption can be written as:

$$Y_t = T_t + S_t + \gamma Y_{t-1} + R_t, \quad (1)$$

with

$$\begin{cases} T_t = \beta_0 + \beta_1 t \\ S_t = \beta_2 \cos(\omega t) + \beta_3 \sin(\omega t) + \beta_4 D_{Sat}(t) + \beta_5 D_{Sun}(t) \end{cases},$$

where calendar time is measured via the cardinality t of the observation, starting from 1 on the first date in the dataset; T_t is the trend term, S_t the seasonality both yearly and weekly, introduced via two dummy variables for Saturday and Sunday, $\omega := 2\pi/365$, and R_t are the residuals; γ and $\{\beta_i\}_{i=0,\dots,5}$ are the regression parameters.

The GLM considers both time effects and an AR term. In order to check the relevance of the AR component when calibrating the model, we proceed following three steps that are standard in time-series econometrics (see, e.g., [6]).⁵ In Sect. 5, we show that an AR term should be included and that an AR(1) describes properly the time-series.

3.2 The GP part

It is well known that, after having detrended and deseasonalized the time-series, the impact on power consumption of weather conditions in general and of temperature in particular, is very important and cannot be neglected (see, e.g., [10]). As already

⁵ First, the parameters related to the time effects are calibrated through an ordinary least squares, including only the trend T_t and the (yearly and weekly) seasonality S_t . Second, we measure the autocorrelation and the partial autocorrelation of the residuals of this regression, verifying the necessity of an AR component in the model. Finally, we refuse the null hypothesis of a unit root via an augmented Dickey-Fuller test.

stated in the introduction, in this study we propose to incorporate weather conditions in the model via a GP, that provides a simple tool for density forecast. In this Section, we briefly recall the main characteristics of GP, using a notation similar to [27].

A GP “is a collection of random variables, any finite number of which have a joint Gaussian distribution” (cf. [27, Def.2.1, p. 13]). It is completely specified by its mean function and covariance matrix. In the case of zero mean, the random variables represent the value of the function $R(x)$ at “location” x ; in [27] they are indicated as:

$$R(x) \sim \mathcal{GP}(0, h(x, \bar{x})) \quad , \quad (2)$$

where $h(x, \bar{x})$ is an arbitrary kernel (or covariance function) between the “locations” x and \bar{x} . In practice, this notation indicates that for any collection of n observations, the corresponding residuals are Gaussian random variables s.t.

$$\mathbf{R} \sim \mathcal{N}(0, H(X, X)),$$

where $\mathcal{N}(\cdot, \cdot)$ is a multinomial Gaussian distribution with zero mean and a positive definite covariance matrix $H \in \mathbb{R}^{n \times n}$, $\mathbf{R} \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times m}$, where m is the number of regressors. In this study, we consider 9 regressors: the 7 weather conditions in the dataset (cf. Sect. 2) and 2 other related to the calendar time t , in order to introduce a yearly calendar effect also in the correlation matrix. In particular, we consider $\cos(\omega t)$ and $\sin(\omega t)$ where $\omega = 2\pi/365$ is defined as in the seasonality S_t in Eq. (1). In the following, we continue to refer to these 9 regressors as weather conditions even if they contain these other two explanatory variables. The covariance H_{ij} between the i th and j th residuals depends on the weather conditions X of the corresponding dates t_i and t_j .

The kernel specifies the covariance between pairs of random variables, e.g. the pair $R(x_i)$ and $R(x_j)$,

$$h(x_i, x_j) := \text{cov}(R(x_i), R(x_j)).$$

An example of covariance function is $h(x_i, x_j) = \sigma^2 \delta_{ij}$, where a positive scalar σ^2 multiplies the Kronecker delta, which is one iff $i = j$ and zero otherwise; in this case, the residuals correspond to Gaussian i.i.d. random variables with variance σ^2 as in the standard linear regression.

In this study, we consider a kernel such that

$$h(x_i, x_j) = k(x_i, x_j) + \sigma^2 \delta_{ij}, \quad (3)$$

where

$$k(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{\|x_i - x_j\|}{\sigma_l}\right), \quad (4)$$

with $\|x_i - x_j\|$ the Standardized Euclidean Distance (SED) between x_i and x_j and $\sigma_f \geq 0$, $\sigma_l > 0$ two additional parameters w.r.t. the standard linear regression (see,

e.g., [17, 21]).⁶ The choice of SED compared to L2 norm allows the same contribution of each weather condition, independently from its unit of measure.

As standard in the statistical literature, the dataset is divided into in-sample (IS) for model calibration (training set X , with n observations) and out-of-sample (OS) for forecasting and for evaluating the quality of the results (test set X_* , with a number of points equal to n_*), respectively 2-years (IS) and 1-year (OS). We compute pairwise, through the chosen kernel, the covariance matrix of a finite number of GP observations

$$H(X, X) = K(X, X) + \sigma^2 I, \quad (5)$$

which depends from the three parameters σ_f , σ_l and σ . The matrix I represents the identity in $\mathbb{R}^{n \times n}$.

A GP presents a great advantage: it is immediate to infer the OS residuals and their distribution. The distribution of the points in the IS and the OS sets is (cf. [27, eq. (2.21), p. 16]):

$$\begin{pmatrix} \mathbf{R} \\ \mathbf{R}_* \end{pmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} K(X, X) + \sigma^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right), \quad (6)$$

where $K(X, X_*)$ denotes the $n \times n_*$ matrix of the covariances evaluated at all pairs of training and test points, and similarly for the other entries $K(X_*, X_*)$ and $K(X_*, X)$. Hereinafter, in order to simplify the notation, we indicate with R_t the residual at time t both IS and OS, where R_t for $t = 1, \dots, n$ are the IS values and R_t for $t = n + 1, \dots, n + n_*$ are the OS ones.

We can use the GP for the prediction of OS residuals. The OS residuals (\mathbf{R}_*) given the IS residuals (\mathbf{R}) and the weather conditions, both IS (X) and OS (X_*), are (cf. [27, eq. (2.22), p. 16]):

$$\mathbf{R}_* | X_*, X, \mathbf{R} \sim \mathcal{N}(\bar{\mathbf{R}}_*, \text{cov}(\mathbf{R}_*, \mathbf{R}_*)),$$

where

$$\bar{\mathbf{R}}_* := E[\mathbf{R}_* | X_*, X, \mathbf{R}] = K(X_*, X) [K(X, X) + \sigma^2 I]^{-1} \mathbf{R}, \quad (7)$$

$$\text{cov}(\mathbf{R}_*, \mathbf{R}_*) = K(X_*, X_*) - K(X_*, X) [K(X, X) + \sigma^2 I]^{-1} K(X, X_*). \quad (8)$$

Plugging the OS residuals into the hybrid model (1), we obtain an ex-post probabilistic forecasting of power consumption. Before describing in detail the ex-post density forecasting technique in the next section (cf. Sect. 4.1), we summarize the main characteristics of the GPX model in Sect. 3.3.

⁶ In the ML literature, a Gaussian process is classified within the so-called non-parametric methods. This name does not indicate the absence of parameters, but the fact that the hypothesis space depends mainly from the dimension of the training set. The Gaussian Process we consider in this study involves three parameters: σ_f , σ_l and σ .

3.3 The hybrid model GPX

The GPX model is obtained by Eq. (1), where the residuals are modeled via a GP. An ex-post point forecasting is straightforward with GPX. Its main strength is that the whole OS (conditional) distribution can be easily obtained. OS log-consumption at day $t = n + i$, given IS consumption and weather conditions up to day t , is a Gaussian r.v. with conditional mean equal to

$$\begin{cases} \bar{Y}_{n+1} = T_{n+1} + S_{n+1} + \gamma Y_n + \bar{R}_{n+1} & i = 1, \\ \bar{Y}_{n+i} = T_{n+i} + S_{n+i} + \gamma \bar{Y}_{n+i-1} + \bar{R}_{n+i} & 2 \leq i \leq n_* \end{cases} \quad (9)$$

where residuals' expected value \bar{R}_{n+1} is obtained in (7) and conditional variance is

$$\begin{cases} \text{var}(Y_{n+1}) = \text{var}(R_{n+1}) & i = 1, \\ \text{var}(Y_{n+i}) = \text{var}(R_{n+i}) + \gamma^2 \text{var}(Y_{n+i-1}) + 2 \sum_{j=1}^{i-1} \gamma^j \text{cov}(R_{n+i}, R_{n+i-j}) & 2 \leq i \leq n_*, \end{cases} \quad (10)$$

where $\text{cov}(R_{n+i}, R_{n+i-j})$ is obtained in (8). Equation (10) is the most relevant modeling result in this paper: it allows to obtain the *ex-post* density forecast (described in Sect. 4.1) for the proposed consumption model (1) and (2). It extends the known formula for autoregressive processes in presence of i.i.d. residuals (see, e.g., [3, Ch.3.2.3, p. 58]) to the case of interest. It can be proven via an induction method.

4 Methodology

This Section delineates the adopted methodology and it is organized as follows: after presenting the ex-post forecasting technique in Sects. 4.1, 4.2 describes the benchmark models and Sect. 4.3 the evaluation methods.

4.1 The ex-post forecasting technique and the flow diagram

The forecast of middle-term daily density power consumption is obtained via an ex-post forecasting: a technique, introduced by [14] in the power consumption sector, now commonly used in middle to long term power consumption forecasts (see, e.g. [9]). As shown in the flow diagram of Fig. 2, the method is divided into three stages (see, e.g., [14, p. 1144]): calibration, forecasting and evaluation.

In the first stage, the GPX model (in its two components GLM and GP) is calibrated with the IS training set, with both power and meteorological data. GLM is calibrated through ordinary least squares (OLS), while the calibration of GP parameters σ_f, σ_l and σ is obtained maximizing the log-likelihood

$$\max_{\sigma_f, \sigma_l, \sigma} \left[-\frac{1}{2} \mathbf{R}^\top H(X, X)^{-1} \mathbf{R} - \frac{1}{2} \ln \det H(X, X) - \frac{n}{2} \log 2\pi \right], \quad (11)$$

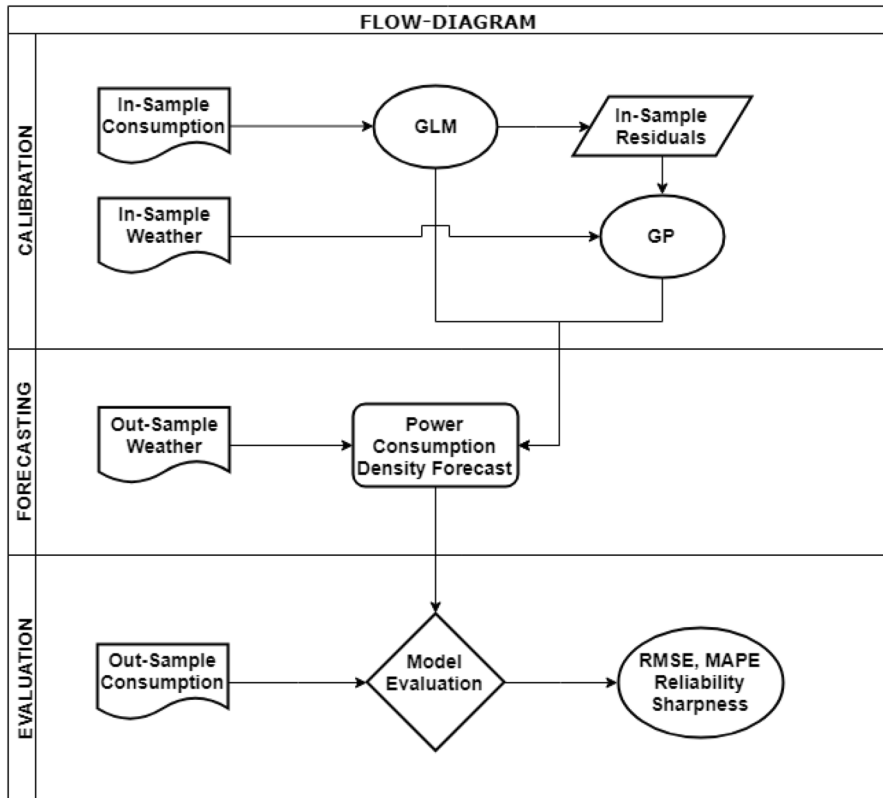


Fig. 2 Flow-diagram of the three stages of the method: calibration, forecasting and evaluation on the proposed model

where $H(X, X)$ is defined in (5). The log-likelihood is maximised through a gradient descent iterative procedure with an adaptive step length (see, e.g., [27], and references therein).

In the second stage, the density forecasting is obtained via an ex-post forecast. This forecast uses the weather conditions in the OS set in order to forecast the power consumption; as well explained by Goude et al. [9, p. 443] “assuming that the realisation of the meteorological covariates is known in advance (...) allows us to quantify the performances of our model without embedding the meteorological forecasting errors”. The idea is that, in order to focus on the ability of the model to describe a strong and reliable relation between daily weather conditions and power consumption, one supposes to know perfectly the weather conditions in the OS period.

Finally, in the third stage, the quality of model forecast is evaluated comparing it with the last year of realized OS consumption data. Model evaluation is realized both in terms of point consumption forecast and of reliability and sharpness of the predicted densities. These evaluation methods are portrayed in Sect. 4.3, after having described the benchmark models.

4.2 Benchmark models

We consider three linear benchmark models in the power industry for a comparative assessment of GPX. As already emphasized in the Introduction, due to the characteristics of the dataset, with only 2-year IS daily data and 1-year OS, benchmark models should be simple and (possibly) parsimonious.

The first one is the Tao benchmark, that was introduced by Hong et al. [12], as part of the process to start up the planning department in a fast growing US utility firm. It has been used as benchmark model for log-consumption in several studies on probabilistic forecast, and it is often referred as Tao's Vanilla benchmark [10, 13, 16]. In this paper, we consider its daily version.

It is well known that there are two seasonal blocks in daily load time-series: week and year. The Tao benchmark, besides a linear trend, considers two classification variables (day-of-the-week and month-of-the-year) with 7 and 12 classes, respectively⁷; it includes as meteorological variable the temperature with a polynomial of order three, and an interaction effect between month and temperature. We observe that the model is a genuine ex-post probabilistic forecast, that relies only on temperature as seasonal driver for power consumption; it does not include any AR term. Unfortunately, the model is not parsimonious relying on 54 explanatory variables.

The second one is GLM. The model has been described in Sect. 3.1 and it is the simplest and most parsimonious ex-ante forecast in this field, because it does not include any meteorological variable; it is equivalent to a GP with a diagonal covariance matrix ($\sigma_f = 0$). This model could clarify whether an autoregressive component can be more relevant than temperature when forecasting daily load time-series.

Finally, we consider an AR exogenous model (ARX), which is a linear model that includes multi-meteorological features as exogenous variables (see, e.g. [3, p.534 et seq.]). It is the simplest (and most parsimonious) extension of GLM that considers, on top of GLM 6 regressors, also the 7 weather regressors described in Table 1, as additional explanatory variables in an ex-post forecast.

For the model evaluation, GPX model is compared to these three benchmarks, not only with the standard evaluation methods for point consumption forecasts, but also with the two most relevant ones for density forecasts. The next Section describes them in detail.

4.3 Evaluation methods

Besides the standard measures of point consumption forecasts as root mean squared error (RMSE) and mean absolute percentage error (MAPE), we provide two evaluation methods of density forecasting. It is more difficult to evaluate a density forecast rather than a point forecast, because we cannot observe the

⁷ For example, the classification variable (or dummy variable) day-of-the-week can be represented as follows by 6 indicator variables (to avoid collinearity): $W_{t,1} = 1$ if the day is Sunday, $W_{t,1} = 0$ otherwise; $W_{t,2} = 1$ if the day is Monday, $W_{t,2} = 0$ otherwise; (...) and $W_{t,6} = 1$ if the day is Friday, $W_{t,6} = 0$ otherwise.

realized distribution of the underlying process. Therefore, we cannot compare the predicted distribution to the true one, as we only have one realisation for each distribution. The evaluation is based on two main measures: the *reliability* that attests distribution’s statistical significance and the *sharpness* that verifies that the forecast is as tight as possible around the expected value. We summarize their main characteristics below.

Reliability is the most important evaluation method for probabilistic forecasting. It refers to the statistical consistency between the probabilistic forecasts and the realized observations OS in the test set. In practice, it determines the fraction of observations that fall outside the confidence interval (CI) with a given nominal level q ; e.g., if the fraction of the realized daily power consumptions, that falls within the 90% CI, is close to 90% then this CI is said to be reliable.

More in detail. Let \hat{L}_t and \hat{U}_t be, respectively, the lower and upper bounds for a given (central) q CI, where q is the CI nominal level, and y_t the actual consumption at time t , the indicator I_t takes two values: 1 if the actual consumption falls within the forecasted CI and zero otherwise, i.e.

$$I_t = \begin{cases} 1 & \text{if } y_t \in [\hat{L}_t, \hat{U}_t] \text{ "hit"} \\ 0 & \text{if } y_t \notin [\hat{L}_t, \hat{U}_t] \text{ "violation"} \end{cases}$$

The empirical coverage is the OS mean of the indicator. Qualitatively, the closer is the empirical coverage to the nominal level, the better it is. In Sect. 5, we show both the nominal level and the empirical coverage for several values of q .

In order to verify that the two sets are close enough even from a quantitative point of view, it has become standard in the banking industry to run two statistical tests. The first one is named *unconditional coverage*: it tests the zero hypothesis that the empirical coverage (i.e. the backtested q CI) equals the nominal level q , and then that $\mathbb{P}(y_t \in [\hat{L}_t, \hat{U}_t]) = q$ [15]. Let n_0 and n_1 be respectively the number of zeros and ones of the indicator I_t , the test is carried out in the likelihood ratio (LR) framework

$$LR_{UC} := -2 \ln \frac{(1 - q)^{n_0} q^{n_1}}{(1 - \pi)^{n_0} \pi^{n_1}}$$

where $\pi = n_1 / (n_0 + n_1)$ is the empirical coverage. LR_{UC} is distributed asymptotically for large n_* as a $\chi^2(1)$ [15].

The second one is named *conditional coverage* and it tests the alternative hypothesis that the ones and the zeros are clustered together in the indicator I_t time-series. In the alternative model, the time-series is modeled as a first-order Markov chain. Let n_{ij} be the number of observations with the value i for the indicator I_t followed by j for I_{t+1} and $\pi_{ij} := n_{ij} / (n_{i0} + n_{i1})$, the LR statistics is

$$LR_{CC} := -2 \ln \frac{(1 - q)^{n_{00} + n_{10}} q^{n_{01} + n_{11}}}{(1 - \pi_{01})^{n_{00}} \pi_{01}^{n_{01}} (1 - \pi_{11})^{n_{10}} \pi_{11}^{n_{11}}}$$

This LR statistics is distributed asymptotically as a $\chi^2(2)$ [4].

Sharpness is measured via the pinball loss function, an error measure for probabilistic forecast that has become popular in the energy sector after GEFCom

competition. Let $\hat{Y}_{t,q}$ be the consumption forecast at the q th quantile, then the pinball loss function can be written as:

$$\text{Pinball}(q) := \frac{1}{n_*} \sum_{t=n+1}^{n+n_*} P(\hat{Y}_{t,q}, q; y_t),$$

where

$$P(\hat{Y}_{t,q}, q; y_t) := \begin{cases} (1-q)(\hat{Y}_{t,q} - y_t), & \text{if } y_t < \hat{Y}_{t,q} \\ q(y_t - \hat{Y}_{t,q}), & \text{if } y_t \geq \hat{Y}_{t,q} \end{cases}.$$

Regarding the pinball loss as a function of the quantile q , not only its value provides a useful information, but also its shape: an asymmetric pinball loss indicates that the density forecast does not reproduce with the same accuracy right and left tails of the true consumption density, whereas a symmetric pinball loss suggests that the shape of the actual distribution is described adequately by the corresponding forecast. We remind that in power consumption both higher consumptions (right tail) and lower ones (left tail) matter, because the latter can lead to negative electricity prices.

In the following section, we summarize the main results in the three stages described in Fig. 2: calibration, forecasting and evaluation.

5 Results

We compare the GPX model with the three benchmarks (Tao, GLM and ARX model). As discussed in previous section (see Fig. 2), the analysis is divided in three stages: calibration, forecasting and evaluation. The four models are calibrated in-sample (2 y data from the 1st of April 2014 to the 31st of March 2016) and the forecasts are evaluated in the out-of-sample set (1 y data from the 1st of April 2016 to the 31st of March 2017).

The *calibration* of the GPX model is implemented considering first its GLM part and then the GP, after a data pre-processing phase.

The data pre-processing consists in the treatment of leap years and outliers. We remove from the dataset February the 29th in leap years. Outliers may influence seasonality analysis. For this reason, they have been removed following the same technique described in Benth et al. [2]. Following that technique, only one outlier has been detected corresponding to the 25th of July 2015. This outlier has been removed and the GLM is calibrated IS. After calibrating the GLM, the outlier has been reinserted into the time-series.

The GLM presents both seasonality and autoregressive parts. In order to include the AR component in the model, we verify the three steps described in Sect. 4 (cf. Footnote 5). First, only the parameters related T_t and S_t in Eq. (1) are calibrated through OLS. Second, we measure the autocorrelation and the partial autocorrelation of regression residuals: Fig. 3 highlights the need of a one day

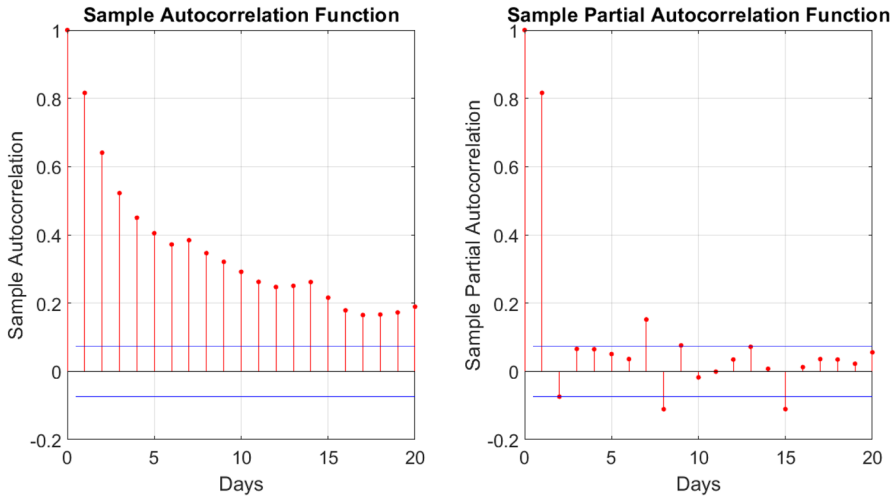


Fig. 3 Autocorrelation function and partial autocorrelation function of seasonally adjusted consumption timeseries. We observe that an AR(1) well explains the observed autocorrelation in the time-series. The horizontal lines indicate the 95% CI

Table 2 GLM parameters calibrated IS with their standard deviation (SE)

		Estimate	(SE)
Intercept	β_0	1.17***	(0.14)
Trend	β_1	- 4.36e-05***	(1.1e-05)
Saturdays	β_2	0.022***	(0.005)
Sundays	β_3	0.047***	(0.005)
Cos	β_4	0.035***	(0.006)
Sin	β_5	-0.103***	(0.012)
AR	γ	0.819***	(0.021)

With *** we indicate statistical significance of the parameters at 1% significance level

Table 3 GP parameters calibrated IS with their standard deviation (SE)

	Estimate	(SE)
σ	3.025e-02***	(6.5e-05)
σ_f	0.240***	(3.8e-03)
σ_l	87.8***	(3.5)

With *** we indicate statistical significance of the parameters at 1% significance level. Standard errors are obtained by means of parametric bootstrapping technique with 10^3 samples

AR component. Third, the augmented Dickey–Fuller test refuses the null hypothesis of a unit root.⁸

⁸ An AR(1) model minimizes both BIC and AIC criteria. Moreover, the model is robust to the inclusion of local and national UK holidays.

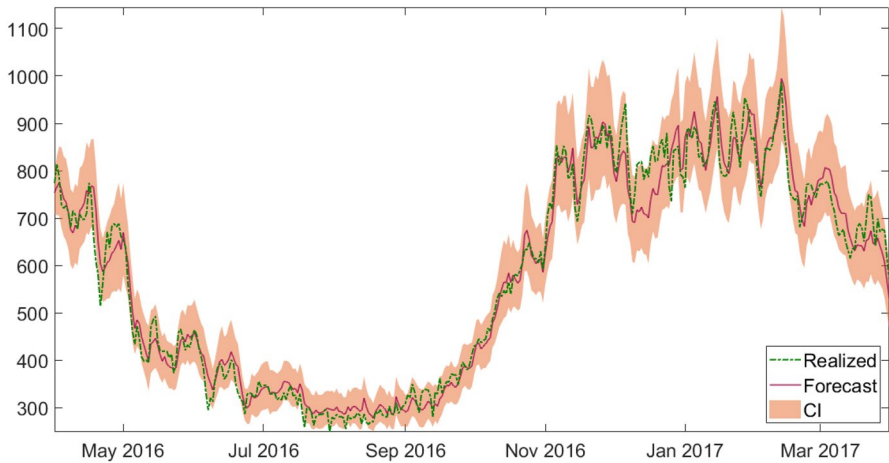


Fig. 4 Realized (dot dashed green line) and expected (continuous pink line) power consumption in MWh between April 2016 to March 2017 with predicted Confidence Intervals at 95% (transparent bright red) (color figure online)

Table 4 RMSE and MAPE for the four models considered

	Tao	GLM	ARX	GPX
RMSE	102.95	79.24	46.92	33.84
MAPE (%)	14.93	9.94	5.90	4.59

Bold values obtained for the proposed model, compared to the values of the benchmark models

We observe that GPX not only presents a MAPE lower than 5%, i.e. it is considered a good forecast by practitioners, but also the lower RMSE indicates a more precise point forecasting in winter times, i.e. when forecasting is more relevant

Then, we perform an IS calibration of all GLM parameters via OLS, reported in Table 2.

The GP part of GPX, described in Sect. 4, is calibrated IS on GLM residuals, maximizing the log-likelihood (11). GP parameters are reported in Table 3; standard errors are obtained via a parametric bootstrapping technique (see, e.g., [5]). All calibrated parameters are statistically significant at 1% significance level.

The *ex-post forecasting* is straightforward with GPX: each density forecast at time $t = n + 1, \dots, t = n + n_*$ is Gaussian with conditional mean (9) and conditional variance (10). In Fig. 4, we show the OS power consumption forecasting of GPX: the continuous pink line indicates the point forecast whereas the transparent bright red indicates the 95% confidence interval; we also show with a dot

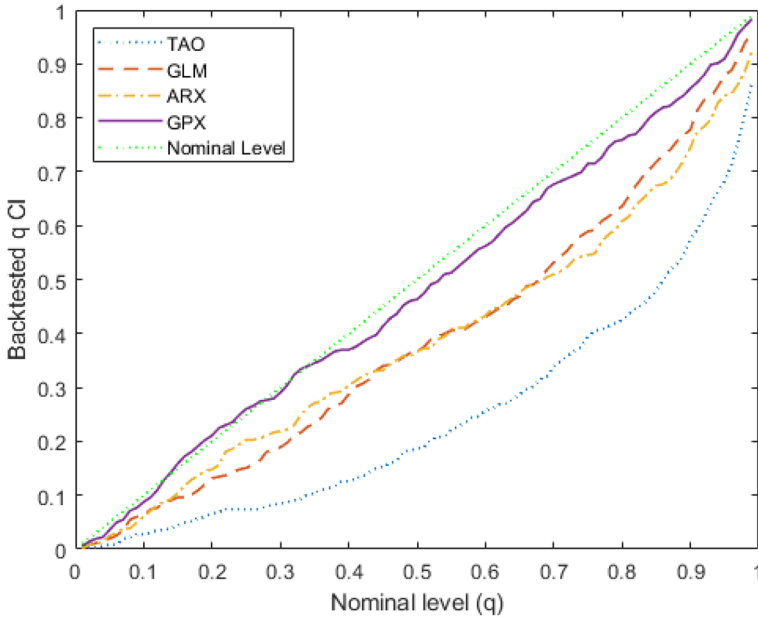


Fig. 5 Empirical coverage. We observe that the backtested q CI for GPX (continuous violet line) are very close to nominal levels q

dashed green line the realized OS power consumption. Results look impressive: the point forecast tracks closely the realized consumption and even the spikes in winter time are tracked very closely. The densities reproduce the observed behavior of periods of low volatility in summer time followed by periods of high volatility in winter time.

Let us underline that the last power consumption considered in model calibration is the 31st of March 2016, whereas the forecast goes up to one year later (31st of March 2017). The quality of this forecast is the main result of the paper.

In the remaining part of this section we provide some quantitative criteria that show the goodness of both point and density forecasting.

The *evaluation* on the model compares GPX with the three benchmark models. We first consider accuracy measures for the point forecasting and then we show the results for the reliability and the sharpness of the density forecasting: these evaluation techniques have been described in Sect. 4.

Table 5 Likelihood ratios tests at 90% level of 99% CI

	Tao	GLM	ARX	GPX	χ^2 statistic
Unconditional coverage	175.25	32.651	39.638	1.280	2.706
Conditional coverage	229.38	88.478	114.738	1.494	4.605

Bold values obtained for the proposed model, compared to the values of the benchmark models

The χ^2 test statistic represents the threshold over which one should reject the null hypothesis

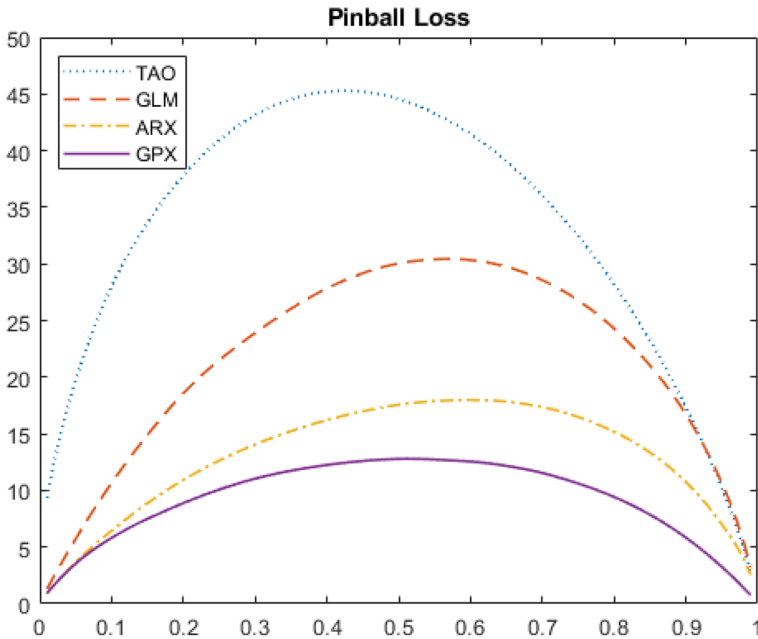


Fig. 6 Pinball loss functions for the four models for the 1st to the 99th percentiles. We observe that not only GPX (continuous violet line) presents the lowest score for all percentiles (i.e. it is sharper and more accurate), but also that the pinball loss is more symmetric for GPX than for all benchmark models

First, we value root mean squared error (RMSE) and mean absolute percentage error (MAPE).

In Table 4, we compare RMSE and MAPE of the four models in the OS set. We observe that GPX and ARX are better than both Tao and GLM, in terms of RMSE and MAPE, being GPX the best one. It is remarkable that the daily version of the Tao model—based only on temperature as weather variable—performs poorly; the other benchmarks suggest that, for a middle-term forecast and a short daily time-series, the AR component is very relevant (cf. GLM) and a multi-meteorological analysis is important (cf. ARX).

In particular, let us emphasize that for GPX the MAPE is lower than 5%, the threshold that limits—for practitioners—a good forecast in daily power consumption. Moreover, a lower RMSE (almost one third lower than ARX) indicates that the GPX reduces significantly the error also in winter times, when the forecast is more relevant due to the higher consumption in absolute terms and the higher volatility: a behavior observed in Fig. 4.

Second, the analysis of reliability is presented. Figure 5 provides the back-tested confidence intervals. The qualitative results of this evaluation method in terms of reliability of the proposed density forecasting look very good: we notice

that GPX backtested q CI are much closer to nominal ones for any choice of nominal level q w.r.t. all benchmarks.

The reliability of our model can be also tested from a quantitative point of view through the two likelihood ratio (LR) tests, presented in Sect. 4 (the *unconditional* and the *conditional coverage*), as shown in Table 5. We underline that only GPX passes the tests, whereas they are rejected by all benchmark models.

Figure 5 together with Table 5 are the strongest results of our evaluation analysis: the GPX model is able to provide reliable confidence intervals over one-year time horizon for daily power consumption. In fact, on the one hand, we have a MAPE lower than 5% over the whole time-horizon, and, on the other hand, the nominal level and the empirical coverage appear very close for all values of q shown in Fig. 5. The reliability is confirmed by the LR tests; the conditional coverage has shown also that the “violations” are not clustered in a particular period of the year, as it is revealed also by a direct inspection in Fig. 4. The results imply that GPX model is able to catch a very accurate relation between weather conditions and power consumption distribution.

Finally, we consider the analysis of sharpness: interesting results arise from the pinball loss. Figure 6 shows pinball loss for the 1st to the 99th percentiles of predictions for GPX and benchmark models. We observe that the plot of the pinball loss provides useful information in terms of sharpness (GPX pinball loss is lower for all percentiles than all benchmark models) and symmetric shape. The proposed density forecasting of power consumption is reproducing with the same accuracy both right and left tails of the actual consumption density.⁹

6 Concluding remarks

This paper introduces the GPX (cf. Eqs. (1) and (2)), a new hybrid model for middle-term probabilistic forecast of power consumption. The model is very parsimonious—with only nine parameters—and can be easily calibrated thanks to its hybrid nature that allows to consider first seasonality and AR characteristics, and then residuals via a Gaussian Process. It has, what’s more, a relatively low computational cost compared to other ML techniques.

Several are the contributions of this study. First, GPX provides a daily load density forecast over a middle-term horizon, a forecast that is very important both for network reliability of power systems and for investment strategies in new plants and transmission facilities. In order to build a very parsimonious model, for middle-term forecast, in the calibration stage we have (i) detailed the construction of the linear part of the model, selecting regressors that are highly significant (cf. Table 2), and (ii) verified the statistical significance of the GP parameters via a parametric bootstrap technique (cf. Table 3). The proposed hybrid modeling approach has improved existing applications of GP to (short-term) power consumption forecasts.

⁹ We have also performed the sharpness analysis in terms of the Winkler score [29], that confirms the results obtained with the pinball loss. These results are available upon request.

Second, the comparison with 3 benchmark models indicates, even with simple metrics, some not-obvious results. Average point estimation is the attempt to provide the single best prediction: MAPE and RMSE are still widely used error statistics in business point forecasting. Even with these point forecasting measures, we have shown in Table 4 that a simple ex-ante model, with only an AR and a seasonal component (GLM), can have a greater forecasting power in daily time-series than an ex-post model, with temperature as the only weather explanatory variable (Tao). Moreover, the knowledge of multi-meteorological conditions improves the forecast and this improvement is more significant when weather variables are included via a Gaussian Process (as in the GPX) rather than linearly (as in an ARX). The GPX produces very accurate forecasts in middle-term forecasting, with a MAPE lower than 5% for daily forecasts over one year (cf. Table 4).

Third, we have shown the relevance of probabilistic forecast of load also in a middle-term time horizon. Load forecasting has become more complicated after the deregulation of the electricity market; thus, an average point prediction is not enough. It is important to understand the distribution of the predicted load; thus, we need to consider even some evaluation methods of the probabilistic uncertainty. Backtesting confidence intervals appears to be the key method when comparing realized and forecast probabilistic consumption. Even a *qualitative* comparison between empirical coverage nominal levels (cf. Fig. 5) is a simple tool with a clear informational content: for linear benchmarks, the fraction of realized loads that falls within the q CI is much lower than the nominal level q , signaling an inadequate probabilistic description, whereas the GPX appears to track closely the theoretical level.

This study highlights backtesting measures as the main evaluation tool for probabilistic load middle-term forecast also from a *quantitative* perspective through two tests: the unconditional and conditional coverage tests. Both of them are elementary to implement and simple to interpret: the unconditional coverage tests the zero hypothesis that the backtested q CI equals the nominal level q , while the conditional coverage verifies that the realizations outside the CI do not cluster. Table 5 shows that GPX passes both tests as opposed to the three benchmarks. In this study, we have improved the existing literature on probabilistic load forecasting, where backtesting quantitative measures are considered rarely.

Finally, we have compared GPX with the benchmark models via the pinball loss, a sharpness measure popular in the energy sector. Results are interesting: Fig. 6 shows, on the one hand, that GPX pinball loss is lower for all percentiles than all benchmarks and, on the other hand, the pinball loss has a symmetric shape for GPX. The latter signals that GPX reproduces with the same accuracy both tails of the distribution of realized power consumption.

This study may have useful applications to a wide range of energy companies. We have observed that a larger set of explanatory variables based on a multi-meteorological dataset is relevant in load forecasting; we have shown that models which take into account several weather information have a significantly better forecasting power, even with short time-series. Unfortunately, when developing in-house load forecasts within a firm, the data quality (especially for weather) may vary; this study can be used by the utility forecasting unit to determine whether or not to install their

own weather stations (to collect meteo data, at least on a daily basis) in some strategic locations within their distribution network. In addition, two of the seven weather conditions considered in this study (cf. e.g. Table 1), wind speed and solar radiation, have a relevant impact also on the production side of renewables.

As for future research, two main promising directions appear evident. First, one key result of this study is that the marginal distribution for daily power consumption is very well approximated by a Gaussian at a regional level. Future work of this paper is to verify whether a hybrid linear-GP model can be a competitive forecasting tool even with time-series longer than the 3-years considered in this paper. Longer time-series would make the comparison possible also with other ML techniques that use the Gaussian distributional property of forecasts, other than linear benchmarks.

Second, an advantage of GPX is that, due to its parsimony, its calibration is accurate with short time-series; in other words, GPX responds quickly to sudden changes of loads. Future work is to implement GPX, in a scenario generation tool at firm level, extending the set of explanatory variables to those relevant during fast macro-economic changes, such as recession or pandemic.

Acknowledgements We thank all participants to the 5th EFI workshop in Rome (2020). We are grateful in particular to F. Cordonì, P. Falbo, C. Lucheroni and P. Sabino for useful comments. We thank M. Azzone and P. Manzoni for fruitful discussions on this topic.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Data availability The data-provider did not give written consent for their data to be shared publicly, so restrictions apply to the availability of these data.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. BCoBS.: Supervisory framework for the use of backtesting in conjunction with the internal models approach to market risk capital requirements. Technical Report Basel Committee on Banking Supervision (1996)
2. Benth, F.E., Benth, J.S., Koekebakker, S.: Stochastic Modelling of Electricity and Related Markets. World Scientific, Singapore (2008)
3. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. Wiley, New York (2015)
4. Christoffersen, P.F.: Evaluating interval forecasts. *Int. Econ. Rev.* **39**, 841–862 (1998)
5. Efron, B., Tibshirani, R.: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1**, 54–75 (1986). <http://www.jstor.org/stable/2245500>
6. Enders, W.: Applied Econometric Time Series, 4th edn. University of Alabama, New York (2015)
7. Fan, S., Chen, L.: Short-term load forecasting based on an adaptive hybrid method. *IEEE Trans. Power Syst.* **21**, 392–401 (2006)

8. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Cambridge (2016)
9. Goude, Y., Nedellec, R., Kong, N.: Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE Trans. Smart Grid* **5**(1), 440–446 (2013)
10. Hong, T., Fan, S.: Probabilistic electric load forecasting: a tutorial review. *Int. J. Forecast.* **32**, 914–938 (2016)
11. Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J.: Probabilistic energy forecasting: global energy forecasting competition 2014 and beyond. *Int. J. Forecast.* **32**, 896–913 (2016)
12. Hong, T., Wang, P., Willis, H.L.: A naïve multiple linear regression benchmark for short term load forecasting. In: 2011 IEEE Power and Energy Society General Meeting, pp. 1–6. IEEE (2011)
13. Hong, T., Wilson, J., Xie, J.: Long term probabilistic load forecasting and normalization with hourly information. *IEEE Trans. Smart Grid* **5**, 456–462 (2013)
14. Hyndman, R., Fan, S.: Density forecasting for long-term peak electricity demand. *IEEE Trans. Power Syst.* **25**, 1142–1153 (2010)
15. Kupiec, P.: Techniques for verifying the accuracy of risk measurement models. *J. Deriv.* **3**, 73–84 (1995)
16. Liu, B., Nowotarski, J., Hong, T., Weron, R.: Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Trans. Smart Grid* **8**, 730–737 (2015)
17. Lourenco, J., Santos, P.: Short-term load forecasting using a Gaussian process model: the influence of a derivative term in the input regressor. *Intell. Decis. Technol.* **6**, 273–281 (2012)
18. Marino, D.L., Amarasinghe, K., Manic, M.: Building energy load forecasting using deep neural networks. In: IECON 2016—42nd Annual Conference of the IEEE Industrial Electronics Society (2016)
19. Meng, K., Dong, Z.Y., Wong, K.P.: Self-adaptive radial basis function neural network for short-term electricity price forecasting. *IET Gener. Transm. Distrib.* **3**, 325–335 (2009)
20. Mocanu, E., Nguyen, H., Gibescu, M., Kling, W.: Deep learning for estimating building energy consumption. *Sustain. Energy Grids Netw.* **6**, 91–99 (2016)
21. Morad, M., Abbas, H.S., Nayel, M., Elbaset, A.A., Galal, A.: Electrical energy consumption forecasting using Gaussian process regression. In: 2018 Twentieth International Middle East Power Systems Conference (MEPCON). IEEE (2018)
22. Mori, H., Ohmi, M.: Probabilistic short-term load forecasting with Gaussian processes. In: Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems. IEEE (2005)
23. Nowotarski, J., Weron, R.: Recent advances in electricity price forecasting: a review of probabilistic forecasting. *Renew. Sustain. Energy Rev.* **81**, 1548–1568 (2018)
24. Osczevski, R.J.: The basis of wind chill. *Arctic* **48**, 372–382 (1995)
25. Park, D.C., El-Sharkawi, M., Marks, R., Atlas, L., Damborg, M.: Electric load forecasting using an artificial neural network. *IEEE Trans. Power Syst.* **6**, 442–449 (1991)
26. Ramakrishna, R., Scaglione, A., Vittal, V., Dall’Anese, E., Bernstein, A.: A model for joint probabilistic forecast of solar photovoltaic power and outdoor temperature. *IEEE Trans. Signal Process.* **67**, 6368–6383 (2019)
27. Rasmussen, C., Williams, C.: *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge (2006)
28. Shi, H., Xu, M., Li, R.: Deep learning for household load forecasting—a novel pooling deep RNN. *IEEE Trans. Smart Grid* **9**, 5271–5280 (2018)
29. Winkler, R.L.: A decision-theoretic approach to interval estimation. *J. Am. Stat. Assoc.* **67**, 187–191 (1972)
30. Xu, L., Hu, M., Fan, C.: Probabilistic electrical load forecasting for buildings using Bayesian deep neural networks. *J. Build. Eng.* **46**, 103853 (2022)
31. Yu, J., Weng, Y., Tan, C.-W., Rajagopal, R.: Probabilistic estimation of the potentials of intervention-based demand side energy management. In: 2015 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 865–870. IEEE (2015)