

State-Space Models for Online Post-Covid Electricity Load Forecasting Competition

JOSEPH DE VILMAREST^{1,2} AND YANNIG GOUDE^{1,2,3}

¹Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, 75006 Paris, France

²Électricité de France Research and Development, 91120 Palaiseau, France

³Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

CORRESPONDING AUTHOR: J. DE VILMAREST (joseph.de-vilmarest@edf.fr)

ABSTRACT We present the winning strategy for the IEEE DataPort Competition on Day-Ahead Electricity Load Forecasting: Post-Covid Paradigm. This competition was organized to design new forecasting methods for unstable periods such as the one starting in Spring 2020. First, we pre-process the data with a statistical correction of the meteorological variables. Second, we apply standard statistical and machine learning models. Third, we rely on state-space models to adapt the aforementioned forecasters. It achieves the right compromise between two extremes. Indeed, machine learning methods allow to learn complex dependence to explanatory variables on a historical data set but fail to forecast non-stationary data accurately. Conversely, purely time-series models such as autoregressives are adaptive in essence but fail to capture dependence to exogenous variables. Finally, we use aggregation of experts, and we leverage the diversity of the set of obtained forecasters to improve our final predictions. The evaluation period of the competition was the occasion of trial and error and we put the focus on the final procedure.

INDEX TERMS Covid-19, electricity load, Kalman filter, forecast, state-space model, time series.

I. INTRODUCTION

ELECTRICITY demand forecasting is a crucial task for grid operators. Indeed the production must balance the consumption as storage capacities are still negligible compared to the load. Time series methods have been applied to address that problem, relying on calendar information and lags of the electricity consumption. Statistical and machine learning models have been designed to use exogenous information such as meteorological forecasts (the load usually depends on the temperature for instance, due to electric heating and cooling).

The field has been thoroughly studied over the past decades as shown in the bibliometric review of [1]. We will not propose here an exhaustive bibliographic study and refer to the recent surveys as [2]–[4]. We focus on recent results in the different forecasting challenges related to this field. The Global Energy Forecasting Competitions (GEFCOM) ([5], [6] and [7]) provide a large benchmark of popular and efficient load forecasting methods. Black box machine learning models such as gradient boosting machines [8] and neural networks [9], [10] rank among the first as well as statistical models like Generalized Additive Models (GAM) [11], [12] or parametric regression models [13], [14].

Ensemble methods or expert aggregation are also a common practice for competitors [15], [16].

The behaviour of the consumption changed abruptly during the coronavirus crisis, especially during lockdowns imposed by many governments. These changes of consumption mode have been challenging for electricity grid operators as historical forecasting procedures performed poorly. Therefore, designing new forecasting strategies to take that evolution into account is important to reduce the cost of forecasting errors and to ensure the stability of the network in the future.

It is to be noted that purely time series methods like autoregressives didn't drift as they are very adaptive in essence. However, they fail to capture the dependence of the load to, for instance, meteorological variables. We claim that state-space models allow the best of both worlds. First, machine learning models trained on historical data are used to design new feature representations. Second, a state-space representation yields a methodology to adapt these complex forecasting models.

Our work extends a previous study on the French electricity load [17] where a state-space approach was presented to adapt generalized additive models in the context of online learning.

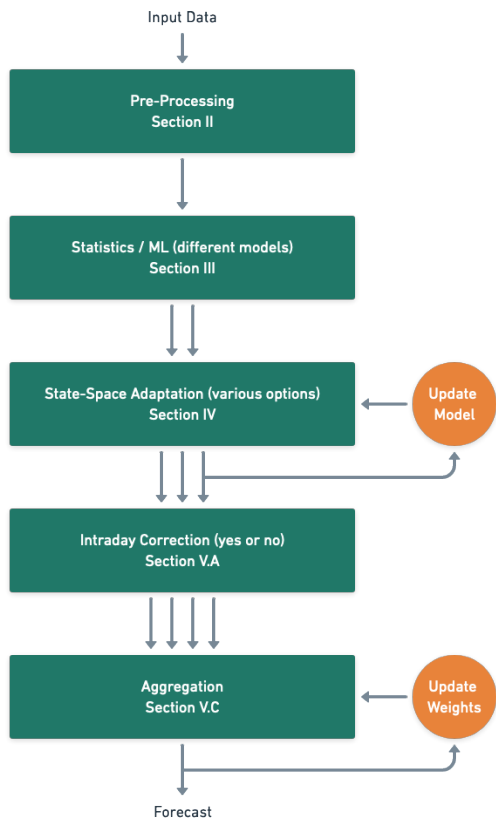


FIGURE 1. Diagram of the forecasting strategy. The multiple arrows mean multiple output. Precisely, the pre-processing step yields one unique output. Then we have several classical forecasting methods, each of which has different adaptation flavors. The intraday correction doubles the number of forecasts, and all of them are combined in the aggregation step, yielding the final forecast.

The idea is to plug a Kalman filter on the estimated effects of a GAM to gain in online reactivity. The novelty of this article lies both in the forecasting method and in the application. First, besides generalized additive models we extend our procedure to other widely used machine learning models including neural networks. Second, after applying the standard Kalman filter [18], we apply another state-space approach named Viking [19]. Viking is a generalization of Kalman filter allowing to estimate jointly the state and the variances in a state-space model. Third, our procedure resulted in the winning strategy in a competition on Post-Covid day-ahead electricity demand forecasting [20], motivating the efficiency of the proposed approach.

A diagram of our forecasting strategy is provided in Figure 1. The article follows its structure. Section II is an introduction to the data set and we detail our pre-processing with a focus on meteorological variables. In Section III we present standard forecasting models. The core of our strategy is Section IV where we propose a generic state-space framework to adapt these methods. We discuss the numerical

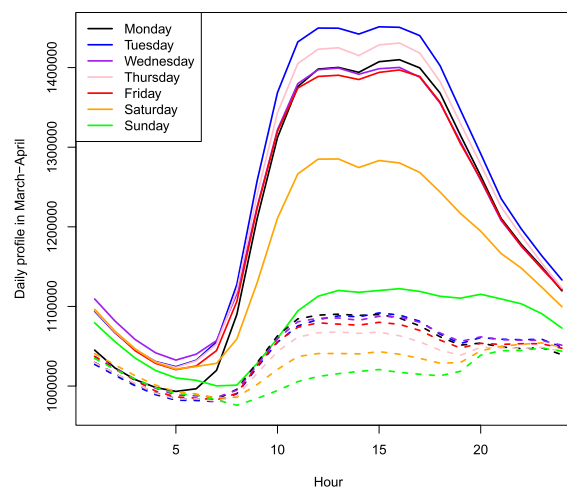
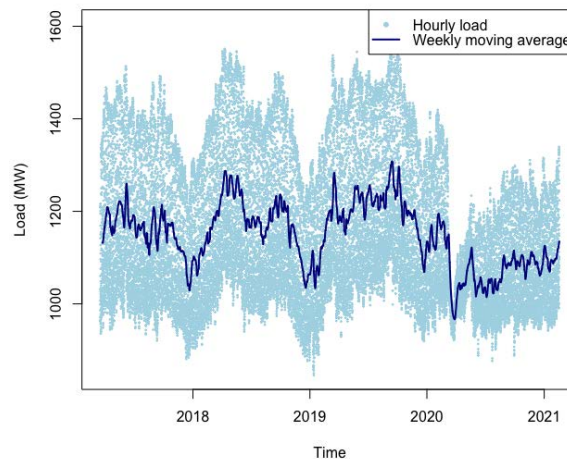


FIGURE 2. On top: electricity load from March 18, 2017, to February 16, 2021. On the bottom: daily profiles of the electricity load in March–April 2019 (solid lines) compared to March–April 2020 (dashed lines).

performances of the various models in Section V and we combine them through aggregation of experts to leverage each model’s advantages.

II. DATA PRESENTATION AND PRE-PROCESSING

The objective of the competition was to predict the electricity load of an undisclosed location of average consumption 1.1 GW, that is of the order of one million people in western countries. The break in the electricity demand in March 2020 is clear in Figure 2. The objective of the competition was to design new strategies for day-ahead forecasting in order to be robust to this unstable period.

A. TIME SEGMENTATION

The competition’s setting was to forecast the hourly load 16 to 40 hours ahead in an online manner. Precisely we had to predict the consumption of each hour of day d with data up to 8AM day $d-1$. After our prediction was sent a new batch of

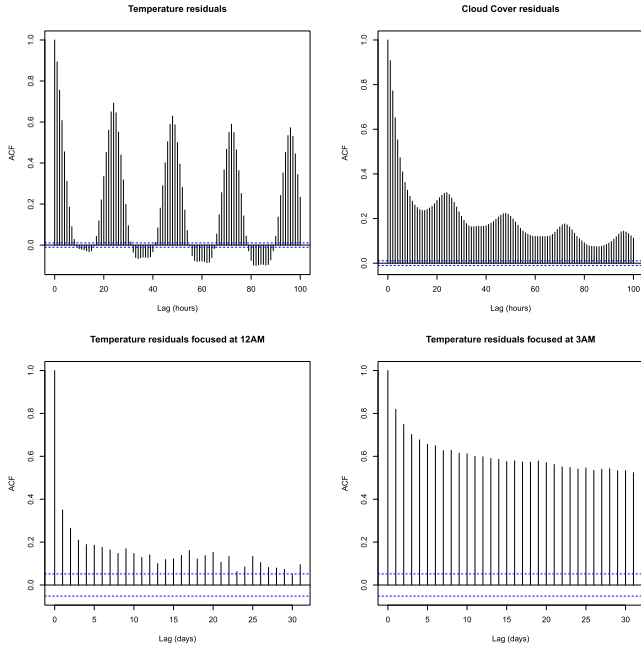


FIGURE 3. On top: auto-correlation plots of the temperature (left) and cloud cover (right) forecasting residuals, with lags in hours. On the bottom: auto-correlation plots of the temperature residuals focused on a specific hour of the day (midnight on the left, 3 a.m. on the right), with lags in days.

data up to 8AM day d was released so that we had to predict day $d+1 \dots$

The evaluation was based on the Mean Average Error on the period ranging from January 18th to February 16th 2021. To build a forecasting model, the historical load starting from March 18th 2017 was provided, as well as meteorological forecasts and realizations during the same period.

B. METEOROLOGICAL FORECASTS

Aside from calendar variables, is usual that the most important exogenous factor explaining the electricity demand is meteorology. The dependence of the load to the temperature for example is due to electric heating and cooling. Moreover, the dependence of the electricity demand to meteorology is augmented by the development of decentralized renewables. Indeed, small renewable production is often used by its owner, yielding a net consumption that highly depends on wind or solar radiation. Therefore the error of a forecasting model for the electricity demand crucially depends on the performance of the meteorological forecasts.

The data of the competition includes forecasts and realization of the temperature, the cloud cover, the pressure, the wind direction and speed. These forecasts are assumed to be known 48 hours in advance and invariant after, thus they can be used to forecast the load at the 16 to 40 hours horizon.

However from the statistical properties of the meteorological forecasting residuals (c.f. Figure 3), we conjecture that the forecasts come from physical models that need to be statistically corrected. Indeed, as the forecasts are available

TABLE 1. Mean Average Error of different meteorological forecasts.

	Initial	Last daily lag	Corrected
Temperature (°C)	3.00	2.11	1.69
Cloud cover (%)	17.28	18.74	14.99
Pressure (kPa)	0.506	1.30	0.423
Wind Speed (km/h)	4.53	3.49	2.53

48 hours in advance, if a statistical correction had been applied then auto-correlations of the residuals over 48 hours would be negligible. We thus use correction models close to autoregressives on the residuals.

Formally, let (z_t) be any of the meteorological variable and (\hat{z}_t) the forecast given in the data set. Then we use the model

$$z_t = \alpha \hat{z}_t + \sum_{l \in \mathcal{L}_{p,P,h(t)}} \beta_l (z_{t-l} - \hat{z}_{t-l}) + \gamma z_{t-l_0(t)} + \delta + \varepsilon_t,$$

where $h(t) \in \{0, \dots, 23\}$ is the hour of the day of time t and

$$l_0(t) = \begin{cases} 24 & \text{if } h \leq 7, \\ 48 & \text{if } h > 7, \end{cases}$$

$$\mathcal{L}_{p,P,h} = \begin{cases} \{24, \dots, 24 * P, h + 17, \dots, h + 16 + p\} & \text{if } h \leq 7, \\ \{48, \dots, 24 * (P + 1), h + 17, \dots, h + 16 + p\} & \text{if } h > 7. \end{cases}$$

In other words, we forecast the residual of the variable of interest with a linear model on

- the last P available daily lags of the residual,
- the last p available lags of the residual (up to 7AM of the previous day),
- the forecast,
- the last daily lag of the variable of interest.

We optimize the coefficients separately for each hour of the day for the temperature, whereas we use the same coefficients at each hour of the day for the cloud cover, pressure and wind speed (except the intercept term). We don't correct the wind direction. The parameters p and P are selected based on BIC. We display in Table 1 the error of the initial forecast, compared to simply using the last daily lag of the variable of interest, and our corrected forecast.

III. TIME-INVARIANT EXPERTS

We summarize the explanatory variables used in our forecasting models:

- calendar variables: the hour of the day, the day of the week, the time of year (Toy) growing linearly from 0 on January 1st to 1 on December 31st, and a variable growing linearly with time to account for a trend,
- meteorological forecasts after statistical correction: the temperature along with exponential smoothing variants of parameters 0.95 and 0.99 (respectively $Temps95$ and $Temps99$), the cloud cover, the pressure, the wind direction and speed,

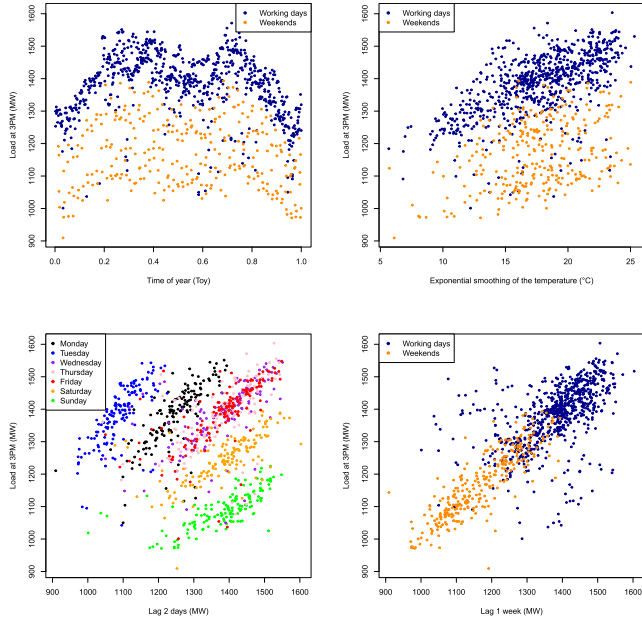


FIGURE 4. Dependence of the load at 3 p.m. to different covariates on the data up to January 1st 2020.

- lags of the electricity load: the load a week ago $LoadW$ and the last load available $LoadD$ (a day ago for the forecast before 8AM and two days ago after 8AM, this constraint coming from the availability of the online data during the competition).

The dependence to the hour of the day and the day of the week is well observed in Figure 2. We display in Figure 4 the dependence of the load to a few of the aforementioned covariates.

We define independent models for the different hours of the day as is usual in electricity load forecasting. For each model we use the same structure for the different hours but we learn the model parameters independently for each time of day based on the training data of that particular time of day. In what follows we denote by y_t the load at time t .

- **Autoregressive.** We consider a seasonal autoregressive model based on the daily and weekly lags of the load:

$$y_t = \sum_{l \in \mathcal{L}_h(t)} \alpha_l y_{t-l} + \sum_{1 \leq l \leq 6} \alpha_{7 \times 24l} y_{t-7 \times 24l} + \varepsilon_t,$$

$$\mathcal{L}_h = \begin{cases} \{24, 48, 72\} & \text{if } h \leq 7, \\ \{48, 72, 96\} & \text{if } h > 7. \end{cases} \quad (1)$$

- **Linear regression.** We use a linear model with the following variables: temperature, cloud cover, pressure, wind direction and speed, day type (7 Booleans), time of year, linear trend variable, and the two lags $LoadW$ and $LoadD$.



FIGURE 5. Evolution of the forecasting error for the different models introduced in Section III trained on the data up to January 1, 2020.

- **Generalized Additive Model (GAM).** We propose a Gaussian generalized additive model [21]:

$$y_t = \sum_{i=1}^6 \beta_i \mathbb{1}_{DayType_t=i} + \gamma Temps95_t + f_1(Toy_t) + f_2(LoadD_t) + f_3(LoadW_t) + \alpha t + \beta_0 + \varepsilon_t,$$

where f_1 is obtained by penalized regression on cubic cyclic splines and f_2, f_3 on cubic regression splines.

- **Random Forest (RF).** We build a random forest [22] with the following covariates: linear trend variable, time of year, day type, the two lags and the two exponential smoothing variables of the temperature. Quantile variants were also computed.
- **Random Forest (RF_GAM).** We also correct the GAM using a random forest on the GAM residuals, with the same covariates as in RF to which we add the GAM effects $f_1(Toy_t), f_2(LoadD_t), f_3(LoadW_t)$ as well as lags (one week, one or two days) of the GAM residuals.
- **Multi-Layer Perceptron (MLP).** Finally we test a multi-layer perceptron of 2 hidden layers of 15 and 10 neurons using hyperbolic tangent activation. We take as input: linear trend variable, time of year, day type, the exponential smoothing variable $Temps95$ and the two lags.

IV. ADAPTATION USING STATE-SPACE MODELS

Due to the lockdowns the consumer's behaviours changed abruptly and therefore the models presented in Section III perform poorly during Spring 2020 and afterwards, see Figure 5. To adapt the models in time, we rely on linear Gaussian state-space models, summarized as

$$\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q_t),$$

$$y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \sigma_t^2),$$

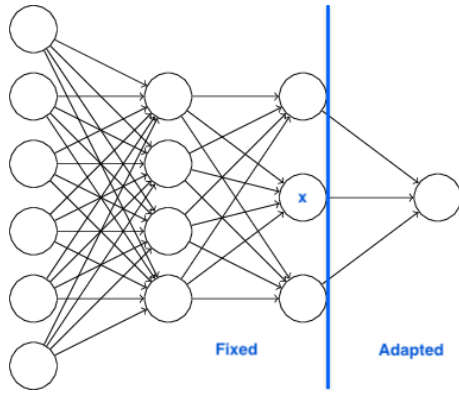


FIGURE 6. Diagram of the definition of the features to adapt the MLP. The network has two hidden layers of 15 and 10 neurons, we freeze all the weights except the last ones.

where θ_t is the latent state, Q_t the process noise covariance matrix and σ_t^2 is the observation variance.

A. DEFINITION OF x_t

This state-space representation is natural for linear regression for which x_t is the vector containing the explanatory variables detailed in Section III. Autoregressive models also fit directly in that framework, as they are in fact linear models on lags of the load, see Equation (1). To adapt GAM and MLP we linearize the models and x_t is just another feature representation. We freeze the non-linear effects in the GAM as in [17], and x_t contains the different effects, linear and non-linear. We apply a similar approach for the MLP, for which we freeze the deepest layers and we learn the last one, that is x_t is the final hidden state, see Figure 6.

The state-space approach is not applied to the random forest. For the latter we compare with incremental offline random forests, consisting in re-training the random forest each day with all the data available at the time.

B. KALMAN FILTER

Bayesian estimation of the state θ_t in linear gaussian state-space models is well understood under known variances σ_t^2, Q_t . The best estimator is obtained by the well known Kalman Filter [18]. It yields recursive exact estimation of the mean and covariance matrix of the state given the past observations, denoted respectively by $\hat{\theta}_t$ and P_t . However there is no consensus in the literature as to how to tune the hyperparameters, see for instance [23]–[25]. The widely used Expectation-Maximization algorithm is an iterative algorithm that guarantees convergence to a local maximum of the likelihood. However there is no global guarantee and in our case it performs poorly. We propose instead the following settings, building on [17]:

- **Static.** We consider the degenerate setting where $Q_t = 0$ and $\hat{\theta}_1 = 0, P_1 = I, \sigma_t^2 = 1$.

- **Static break.** We consider a break at March 1st 2020 by setting $\hat{\theta}_1 = 0, P_1 = I, \sigma_t^2 = 1, Q_t = 0$ except $Q_T = I$ where T is March 1st 2020.
- **Dynamic.** We approximate the maximum-likelihood for constant σ_t^2, Q_t . We set $P_1 = \sigma^2 I$ and we observe that for a given Q/σ^2 we have closed-form solutions for $\hat{\theta}_1, \sigma^2$. Then we restrict ourselves to diagonal matrices Q/σ^2 whose nonzero coefficients are in $\{2^j, -30 \leq j \leq 0\}$ and we apply a greedy procedure: starting from $Q/\sigma^2 = 0$ we change at each step the coefficient improving the most the likelihood. That procedure is designed to optimize Q on the training data (up to January 1st 2020).
- **Dynamic break.** We use similar $\hat{\theta}_1, P_1, \sigma_t^2 = \sigma^2, Q_t = Q$ as in the dynamic setting except $Q_T = P_1 = \sigma^2 I$ where T is March 1st 2020.
- **Dynamic big.** We simply use $\sigma^2 = 1$ and a matrix Q proportional to I defined based on the 2020 data.

Also, it is important that we estimate a gaussian posterior distribution, therefore we have a probabilistic forecast for the load. Precisely, our estimate is $\theta_t \sim \mathcal{N}(\hat{\theta}_t, P_t)$, thus we have $y_t \sim \mathcal{N}(\hat{\theta}_t^\top x_t, \sigma^2 + x_t^\top P_t x_t)$. The likelihood that is optimized to obtain the dynamic setting is built on that probabilistic forecast of y_t given the past observations. In the competition we added quantiles of these gaussian distributions as forecasters in the expert aggregation.

C. DYNAMICAL VARIANCES

The idea behind the break settings introduced in the previous paragraph is that we would like the model to adapt faster during an evolving period such as a lockdown than before. However it consists in modelling a break in the data, a sudden change of state resulting from a noise of much bigger variance at a specific time specified *a priori*. A way to extend the approach would be to define a time-varying covariance matrix depending for instance of a lockdown stringency index such as defined by [26]. However the competition policy forbade the use of external data and the location was undisclosed.

In a more long-term perspective let it be hoped that the evolution of the electricity load won't be driven by lockdowns. It is therefore more generic to learn the variances of the state-space model in an adaptive fashion. We therefore apply a novel approach for time-series forecasting introduced in [19], and named Variational Bayesian Variance Tracking, alias Viking. We briefly recall how the method works. This method was designed in parallel of the competition and was improved afterwards. We present the last version only.

We treat the variances as latent variables and we augment the state-space model:

$$\begin{aligned} a_t - a_{t-1} &\sim \mathcal{N}(0, \rho_a), & b_t - b_{t-1} &\sim \mathcal{N}(0, \rho_b), \\ \theta_t - \theta_{t-1} &\sim \mathcal{N}(0, \exp(b_t)I), \\ y_t - \theta_t^\top x_t &\sim \mathcal{N}(0, \exp(a_t)). \end{aligned}$$

Instead of estimating the state θ_t with variances fixed *a priori*, we estimate both the state and the variances represented

by a_t, b_t . Although we have removed σ_t^2, Q_t as hyperparameters, we now have to set priors on a_0, b_0 along with the parameters ρ_a, ρ_b controlling the smoothness of the dynamics on the variances.

We apply a Bayesian approach. At each step, we start from a prior $p(\theta_{t-1}, a_{t-1}, b_{t-1} \mid \mathcal{F}_{t-1})$ obtained at the last iteration, where we introduce the filtration of the past observations $\mathcal{F}_t = \sigma(x_1, y_1, \dots, x_{t-1}, y_{t-1})$. Then we obtain a prediction step thanks to the dynamical equations yielding $p(\theta_t, a_t, b_t \mid \mathcal{F}_{t-1})$. Finally Bayes' rule yields the posterior distribution $p(\theta_t, a_t, b_t \mid \mathcal{F}_t)$.

However the posterior distribution is analytically intractable, therefore the principle of Viking is to apply the classical Variational Bayesian approach [27]. The posterior distribution is recursively approximated with a factorized distribution. In our setting we look for the best product $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})$ approximating $p(\theta_t, a_t, b_t \mid \mathcal{F}_t)$. The criterion minimized is the Kullback-Leibler (KL) divergence

$$KL(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \parallel p(\theta_t, a_t, b_t \mid \mathcal{F}_t)),$$

where $KL(p, q) = \int \log(dp/dq)dp$. At each step it yields a coupled optimization problem in the three gaussian distributions. The classical iterative method (see for instance [28]) consists in computing alternately $\exp(\mathbb{E}[\log p(\theta_t, a_t, b_t \mid \mathcal{F}_t)])$ where the expected value is taken with respect to two of the three latent variables, and identifying the desired first two moments with respect to the other latent variable. However the expression $\exp(\mathbb{E}_{\theta_t, b_t}[\log p(\theta_t, a_t, b_t \mid \mathcal{F}_t)])$ doesn't match a gaussian distribution in a_t , and similarly for b_t . We therefore use the first two moments of the gaussian distribution to derive an upper-bound of the KL divergence for which we have an analytical solution. We refer to [19] for the detailed derivation of the algorithm.

V. EXPERIMENTS

We display the performance of the introduced methods that we call experts. Then we use aggregation of experts to leverage specificities of each forecaster. The end of the section is devoted to a discussion of our day-to-day strategy during the competition. Finally, we refer to the implementation for more details.¹

A. INTRADAY CORRECTION

Although it performs better to use different models at the different hours of the day, let it be noted that the correlation between different hours is important. To capture intraday information we fit on the residuals of each model an autoregressive model incorporating lags of the 24 last available hours and optimized for each forecast horizon. This follows from the intuition that to predict the load at 8AM, instead of using as the last available data a delay of 48 hours, we can use a 25-hour delay.

¹<https://gitlab.com/JosephdeVilmarest/state-space-post-covid-forecasting>

TABLE 2. Mean Average Error (in MW) of each method of Section II during normal test period.

Adaptation	AR	Linear	GAM	RF	RF_GAM	MLP
No intraday	29.3	20.8	20.7	24.6	23.0	21.2
With intraday	27.0	19.9	19.3	24.4	23.7	20.6

Models are trained up to Jan. 1st 2020 and tested during the next two months before the break of March.

TABLE 3. Mean Average Error (in MW) of each method during the competition evaluation set (2021-01-18 to 2021-02-16).

Adaptation	AR	Linear	GAM	MLP
Offline	14.6	22.8	22.7	16.7
Static	20.5	15.7	17.0	22.9
Static break	27.9	14.4	28.4	35.4
Dynamic	14.4	14.9	15.3	13.0
Dynamic break	16.2	13.6	14.3	12.3
Dynamic big	14.3	11.2	12.4	12.4
Viking	14.4	11.5	12.7	12.5

The performances are displayed for each model after intraday correction. As a comparison, re-training the random forest every day yields an online RF of MAE 15.0 MW, and an online GAM_RF of MAE 18.1 MW. The organizers propose a naive benchmark (relying on persistence) of MAE 15.5 MW.

We apply this correction to the models presented in Section III as well as to the ones resulting from the adaptation framework of Section IV. We display in Table 2 the gain for the statistical and machine learning methods of Section III. To present the improvement brought by the intraday correction we give the performance during a stable period (after the training of the model, but before the Covid crisis). We observe that the only model for which the intraday correction doesn't improve the performance (RF_GAM) is the one including already a residual correction. The improvement during the evaluation period (2021) is much bigger (57% decrease of the MAE for the MLP for instance), and it is natural as the intraday correction is an autoregressive, that is, an adaptive model.

B. ADAPTATION OF INDIVIDUAL EXPERTS

Then we focus on adaptive models to show the improvements due to each setting, see Table 3. We have 4 different models (autoregressive, linear, GAM and MLP). For each one, we try the various adaptation settings (no adaptation, Kalman filters and Viking). Kalman filters with constant covariance matrix proportional to the identity obtain the best results. That is not the case on the data previous to the competition and it depends on the intrinsic evolution of the data.

We illustrate the different settings in Figure 7 where we display the evolution of the state coefficient for the GAM adaptation strategies.

Furthermore, in Figure 8 we present the evolution of the GAM model adapted by Viking for the 24 different hours of the day. This as well as Figure 2 shows the necessity of different models for the different hours of the day. However, the resemblance of close hours motivates the intraday correction to benefit from the correlation between hours, see Section V-A.

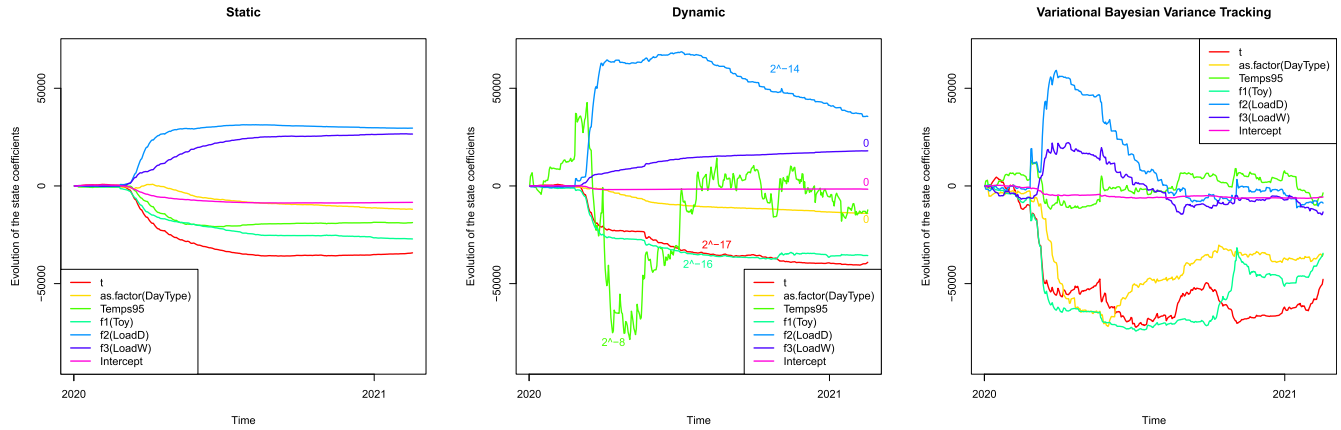


FIGURE 7. Evolution of the state coefficients for various adaptations of the GAM, see Section IV. On the left, we use the Kalman Filter in the static setting (degenerate covariance matrix $Q_t = 0$). On the middle, the dynamic setting where the variances are constant, and we provide the ratio $Q/\sigma^2 = \text{diag}(2^{-17}, 0, 2^{-8}, 2^{-16}, 2^{-14}, 0, 0)$: we observe that the coefficient corresponding to the biggest coefficient of Q (the effect of *Temps95*) evolves much faster. On the right, the Viking setting where we estimate the variances adaptively.

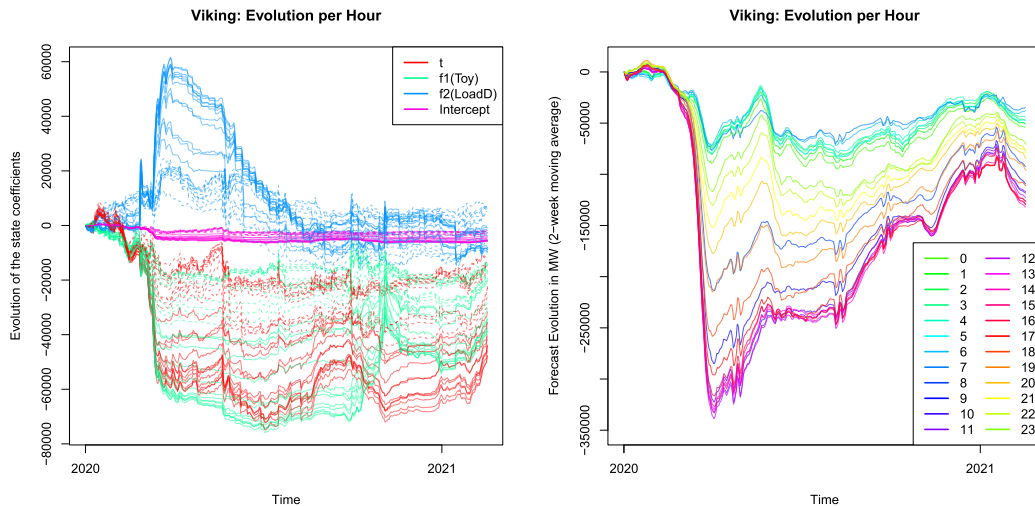


FIGURE 8. Evolution of the Viking adaptation of the GAM for the 24 different hours of the day. On the left: evolution of 4 of the 7 coefficients as in Figure 7. To show the groups we plot night hours (8 p.m.–8 a.m.) in dashed lines and daylight hours (8 a.m.–8 p.m.) in solid lines. On the right: impact of the evolution on the forecast. Precisely on the right graph we display the difference between the forecast of Viking and the forecast that would have been made by Viking on January 1, 2020.

C. AGGREGATION

Online robust aggregation of experts [29] is a powerful model agnostic approach for time series forecasting, already applied to load forecasting during the lockdown in [17]. We use the ML-Poly algorithm proposed in [30] and implemented in the R package *opera* [31] to compute these online weights.

The aggregation weights are estimated independently for each hour of the day. We summarize different variants in Table 4. First, for each family of models we compute the aggregation of all the adaptation settings (7 for each). Then we aggregate all of them (28 models). An example of the weights obtained at 3PM is displayed in Figure 9. The aggregation presented in this paper obtains a performance close to our strategy winning the competition (degradation of about 0.05 MW).

TABLE 4. Mean average error of aggregation strategies (in MW) during the competition evaluation set (2021-01-18 to 2021-02-16).

Adaptation	AR	Linear	GAM	MLP	All
Best expert	14.3	11.2	12.7	12.3	11.2
Aggregation	14.4	11.4	11.7	11.9	10.9

D. DAY-TO-DAY FORECASTS

During the competition our predictions were not exactly the ones of the aggregation method presented in the previous subsection. There are mainly two reasons for that.

First, we considered a bigger set of forecasting methods (we had 72 experts). It seemed reasonable to prune the strategy for the sake of clarity of paper at the cost of a very small change of error, but it is interesting to present also the

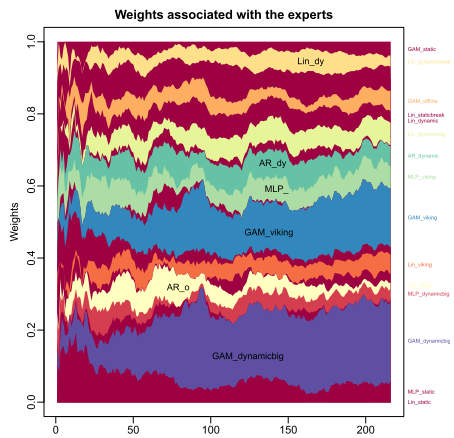


FIGURE 9. Evolution of the aggregation weights at 3 p.m. from July 1, 2020 to February 16, 2021.

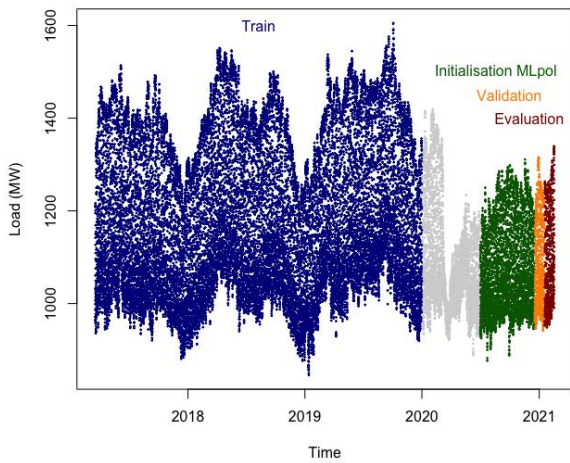


FIGURE 10. Segmentation of the data set. Meteorological corrections as well as time-invariant forecasts were trained on the train period (up to January 1, 2020). Adaptive forecasting methods were evolving on the whole period. Then the aggregation weights were trained by MLpol from July 1, 2020, and the expert selection was determined with respect to the validation set.

predictions used during the evaluation. We found a trade-off in the selection of experts. Indeed too much experts in the aggregation yields poor performances. We applied a greedy procedure to select the experts we keep in the aggregation: we begin with an empty set, and at each step we add the one improving the most the performance. That performance was evaluated with the MAE on the last month of the training data set. We provide in Figure 10 a graphical representation of how we defined different time periods. We refer to Figure 11 for the evolution of the validation MAE as the selection grows. We observe a sharp decrease of the error as experts are added with high diversity and then a slow increase of the loss as the set of experts becomes too large.

Second, we were constantly experimenting the different strategies. We used a variational Bayesian method that was

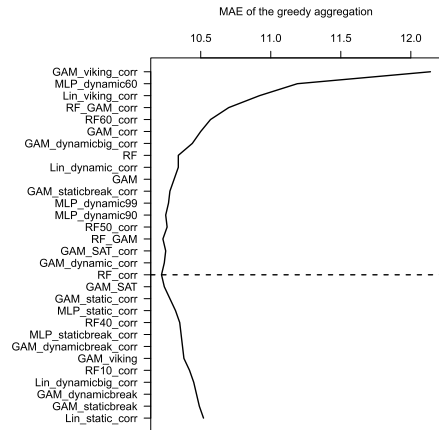


FIGURE 11. Evolution of the validation MAE as the expert selection grows from 1 to 30 experts. The nomenclature is provided in the appendix.

a prior version of the one of this paper. We also changed a lot the aggregation procedure.

We refer to the appendix for a detailed presentation of our daily strategy. Official results of the challenge and additional significance analysis can be found in [32]. Overall these day-by-day changes degraded the performance, if we had stayed on the first strategy with no change at all, our MAE would have been 10.51 MW instead of 10.84 MW. The critical issue in such unstable periods is to find the right validation period to select the prediction procedure. The month before the evaluation period seems *a posteriori* a good compromise. During the competition we changed “manually” based on the performances in a shorter range, for instance considering an expert performing well on the last few weeks for a specific day type ... We should have trusted the aggregation’s robustness.

VI. CONCLUSION

In this paper we presented our procedure to win a competition on electricity load forecasting during an unstable period. Our approach relies heavily on state-space models and the competition was the first data set on which was applied a recent approach to adapt the variances of a state-space model. Some perspectives have been raised during the competition such as interpretability of the global approach and a better understanding of the error propagation along the different adaptations (intraday correction, Kalman filtering, variance tracking and aggregation).

Finally, similar state-space methods have been applied to obtain the first place in another competition in which the objective was to forecast the electricity consumption of a building.²

APPENDIX A. NOMENCLATURE

The experts AR, Lin, GAM, RF, RF_GAM, MLP are the ones presented in that same order in Section III.

²<http://www.gecad.isepp.pt/smartgridcompetitions/>

Names of the form `model_setting` refer to the expert obtained by state-space adaptation of the model `model` with the setting `setting`. For instance, `Lin_dynamic` refers to a linear model adapted with the Kalman filter in the dynamic setting, *c.f.* Section IV-B.

We consider quantile variants of RF, denoted by `RF α` where α is the quantile order in percent (*e.g.* `RF40` is the quantile random forest of quantile value 0.4). We also consider a quantile variant of the dynamic MLP denoted by similar names (`MLP_dynamic60` is the quantile 0.6 of the MLP in the dynamic setting).

Furthermore, we introduce an expert named `GAM_SAT` forecasting each day with the GAM as if it were a Saturday motivated by [17].

Finally, each expert `x` yields another expert `x_corr` after intraday correction.

B. DAY-TO-DAY EVOLUTION OF THE FORECASTING STRATEGY

As explained in Section V-D, our strategy evolved in time and we recall here every change.

- **From January 18th to January 24th:** we used the following set of experts obtained by the greedy selection described in Section V-D: `RF`, `RF_corr`, `RF50_corr`, `RF60_corr`, `Lin_dynamic_corr`, `Lin_viking_corr`, `GAM`, `GAM_corr`, `GAM_staticbreak_corr`, `GAM_dynamic_corr`, `GAM_viking_corr`, `GAM_dynamicbig_corr`, `RF_GAM`, `RF_GAM_corr`, `GAM_SAT_corr`, `MLP_dynamic60`, `MLP_dynamic90`, `MLP_dynamic99`. We aggregated with ML-poly with an aggregation estimated independently for each hour, with the absolute loss. We found afterwards a bug in `RF50_corr`, `RF60_corr`: the quantile RF were set to 0 on the test set so that these two experts were simple intraday autoregressive trained in an unwanted manner.
- **From January 25th to January 31st:** we removed the experts `RF50_corr`, `RF60_corr` and we replaced them with `AR_corr`.
- **February 1st and 2nd:** we used the uniform average between three forecasts. First, the previous aggregation. Second, another aggregation procedure called RF-stacking, consisting in a quantile random forest minimizing the MAE and taking as input the 72 experts as well as the day type and hour of the day. Third, a benchmark close to the one given by the competition organizers: we predict each time with the last available load of the same hour and the same day group (week days, Saturdays and Sundays).
- **From February 3rd to 7th:** we removed the benchmark which damaged the performances. For **Feb. 5th** we corrected the ML-poly prediction using a special day correction, once we observe that Feb. 5th had a special behaviour in the last three years. Precisely, we observed that the relative error of the model is significantly

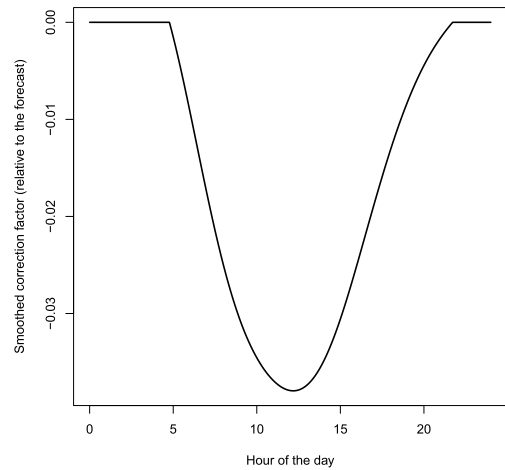


FIGURE 12. Special day correction applied on February 5. It is a multiplicative correction, *e.g.* at midday we reduce our forecast by about 3.8%.

negative on the last three years, a behaviour that may come from a bank holiday for instance. Therefore we fit a smoothed function of the time of day on the relative error and we applied it to our forecast. We truncated so that there is no correction during night. See the shape of the correction in Figure 12.

- **February 8th:** we used the single expert `Lin_dynamicbig_corr` as we observed that it was by far our best expert on the last week, and it seemed to perform especially well on Mondays.
- **February 9th:** we came back to the average between ML-poly and the RF-stacking but we added to the aggregation ML-poly the expert `Lin_dynamicbig_corr`, and we replaced the expert `AR_corr` with another expert `AR_intra` incorporating directly the intraday correction in the autoregressive, instead of correcting an autoregressive based only on daily lags.
- **February 10th and 11th:** we removed the RF-stacking which degraded our performances since its introduction and we kept only the aggregation ML-poly.
- **February 12th and 13th:** we corrected *a posteriori* the electricity load for February 5th with the special day correction. It was important to do it on that day as the weekly lags is important in the models.
- **February 14th:** we used once again the average between the ML-poly aggregation and the RF-stacking, as we observed that the RF-stacking is especially good on Sunday.
- **February 15th and 16th:** we used only the ML-poly aggregation.

REFERENCES

- [1] D. Yang, J.-E. Guo, J. Li, S. Wang, and S. Sun, "Knowledge mapping in electricity demand forecasting: A scientometric insight," *Frontiers Energy Res.*, vol. 9, p. 633, Oct. 2021.
- [2] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecasting*, vol. 32, no. 3, pp. 914–938, 2016.

- [3] A. Almalaq and G. Edwards, "A review of deep learning methods applied on load forecasting," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 511–516.
- [4] I. K. Nti, M. Teimeh, O. Nyarko-Boateng, and A. F. Adekoya, "Electricity load forecasting: A systematic review," *J. Electr. Syst. Inf. Technol.*, vol. 7, no. 1, pp. 1–19, Dec. 2020.
- [5] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *Int. J. Forecasting*, vol. 30, no. 2, pp. 357–363, Apr./Jun. 2014.
- [6] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *Int. J. Forecasting*, vol. 32, no. 3, pp. 896–913, 2016.
- [7] T. Hong, J. Xie, and J. Black, "Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting," *Int. J. Forecasting*, vol. 35, no. 4, pp. 1389–1399, 2019.
- [8] J. R. Lloyd, "GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes," *Int. J. Forecasting*, vol. 30, no. 2, pp. 369–374, Apr./Jun. 2014.
- [9] S. Ryu, J. Noh, and H. Kim, "Deep neural network based demand side short term load forecasting," *Energies*, vol. 10, no. 1, p. 3, Dec. 2016.
- [10] I. Dimoulkas, P. Mazidi, and L. Herre, "Neural networks for GEFCom2017 probabilistic load forecasting," *Int. J. Forecasting*, vol. 35, no. 4, pp. 1409–1423, Oct. 2019.
- [11] R. Nedellec, J. Cugliari, and Y. Goude, "GEFCom2012: Electric load forecasting and backcasting with semi-parametric models," *Int. J. Forecasting*, vol. 30, no. 2, pp. 375–381, Apr. 2014.
- [12] V. Dordonnat, A. Pichavant, and A. Pierrot, "GEFCom2014 probabilistic electric load forecasting using time series and semi-parametric regression models," *Int. J. Forecasting*, vol. 32, no. 3, pp. 1005–1011, 2016.
- [13] N. Charlton and C. Singleton, "A refined parametric model for short term load forecasting," *Int. J. Forecasting*, vol. 30, no. 2, pp. 364–368, Apr./Jun. 2014.
- [14] F. Ziel, "Quantile regression for the qualifying match of GEFCom2017 probabilistic load forecasting," *Int. J. Forecasting*, vol. 35, no. 4, pp. 1400–1408, Oct. 2019.
- [15] P. Gaillard, Y. Goude, and R. Nedellec, "Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting," *Int. J. Forecasting*, vol. 32, no. 3, pp. 1038–1050, 2016.
- [16] S. Smyl and N. G. Hua, "Machine learning methods for GEFCom2017 probabilistic load forecasting," *Int. J. Forecasting*, vol. 35, no. 4, pp. 1424–1431, Oct. 2019.
- [17] D. Obst, J. de Vilmarest, and Y. Goude, "Adaptive methods for short-term electricity load forecasting during COVID-19 lockdown in France," *IEEE Trans. Power Syst.*, vol. 36, no. 5, pp. 4754–4763, Sep. 2021.
- [18] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *J. Basic Eng.*, vol. 83, no. 1, pp. 95–108, 1961.
- [19] J. de Vilmarest, Y. Goude, and O. Wintenberger, *Viking: Variational Bayesian Variance Tracking Winning a Post-COVID Day-Ahead Electricity Load Forecasting Competition*. [Online]. Available: https://roseyu.com/time-series-workshop/submissions/2021/TSW-ICML2021_paper_15.pdf
- [20] M. Farrokhhabadi, *Day-Ahead Electricity Demand Forecasting: Post-COVID Paradigm*, 2020, doi: [10.21227/67vy-bs34](https://doi.org/10.21227/67vy-bs34).
- [21] S. N. Wood, *Generalized Additive Models: An Introduction With R*. Boca Raton, FL, USA: CRC Press, 2017.
- [22] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [23] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. New York, NY, USA: Springer-Verlag, 2016.
- [24] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*. Oxford, U.K.: Oxford Univ. Press, 2012.
- [25] L. Fahrmeir, "Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models," *J. Amer. Stat. Assoc.*, vol. 87, no. 418, pp. 501–509, Jun. 1992.
- [26] T. Hale *et al.*, "A global panel database of pandemic policies (Oxford COVID-19 government response tracker)," *Nature Hum. Behav.*, vol. 5, no. 4, pp. 529–538, Apr. 2021.
- [27] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Springer, 2006.
- [28] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, Jan. 2008.
- [29] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge Univ. Press, 2006.
- [30] P. Gaillard, G. Stoltz, and T. Van Erven, "A second-order bound with excess losses," in *Proc. Conf. Learn. Theory*, 2014, pp. 176–196.
- [31] P. Gaillard and Y. Goude. (2016). *Opera: Online Prediction by Expert Aggregation*. [Online]. Available: <https://CRAN.R-project.org/package=opera.rpackage> version
- [32] M. Farrokhhabadi, J. Browell, Y. Wang, W. Makonin, and H. Zareipour, "Day-ahead electricity demand forecasting competition: Post-COVID paradigm," submitted for publication.



JOSEPH DE VILMAREST graduated from the École Normale Supérieure de Paris in 2019. He is currently pursuing the joint Ph.D. degree with the Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, and Électricité de France (EDF) Research and Development. His research interests include state space models, Bayesian methods, and time series forecasting.



YANNIG GOUDE received the Ph.D. degree in statistics and probability from Université Paris-Sud, Orsay, France, in 2008. He is a Senior Researcher in the field of data science at Électricité de France (EDF) Research and Development and an Associate Professor at Université Paris-Sud. His research interests are time series forecasting for electricity markets, time series analysis, non-parametric models, and aggregation of experts.