Probabilistic Energy Forecasting: Methodologies, Implementations, and Applications

IEEE Power & Energy Society General Meeting
Sunday 17 July 2022
Denver, Colorado

Instructors: Jethro Browell, Tao Hong, Yi Wang, and Bri-Mathias Hodge

Probabilistic Energy Forecasting: Methodologies, Implementations, and Applications

Part 3: State-of-the-art methods from statistical and machine learning

Instructor: Dr Yi Wang

The University of Hong Kong

Four main topics:

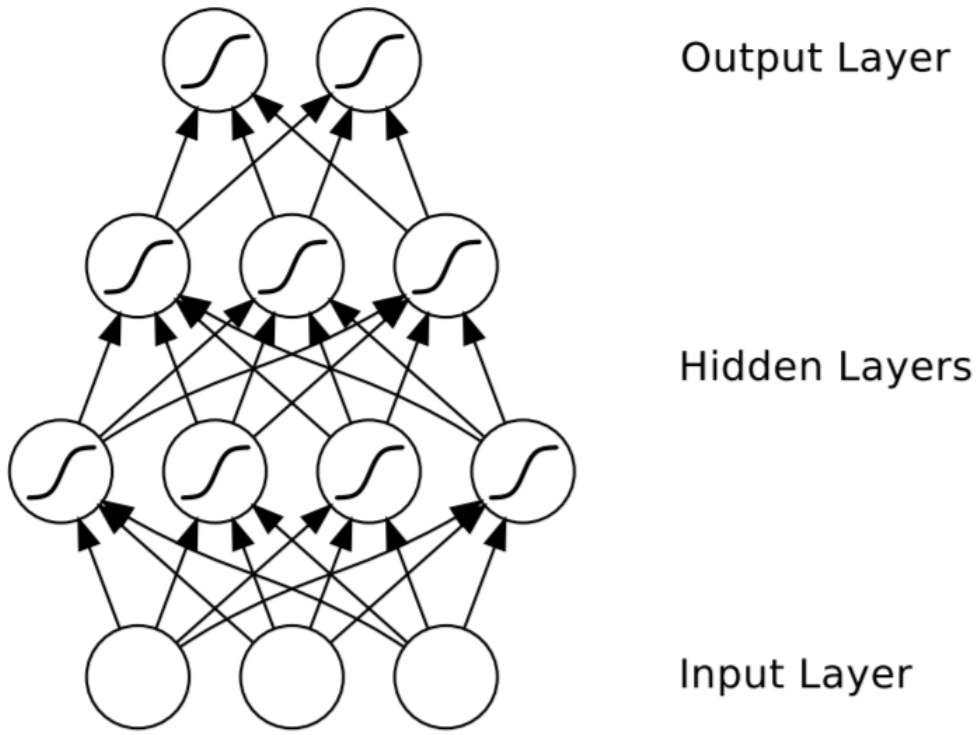
- Deep learning for probabilistic forecasting
- Combining probabilistic forecasts
- Data sharing for probabilistic forecasting

Deep learning for probabilistic forecasting

Deep learning for probabilistic forecasting

Artificial Neural Network

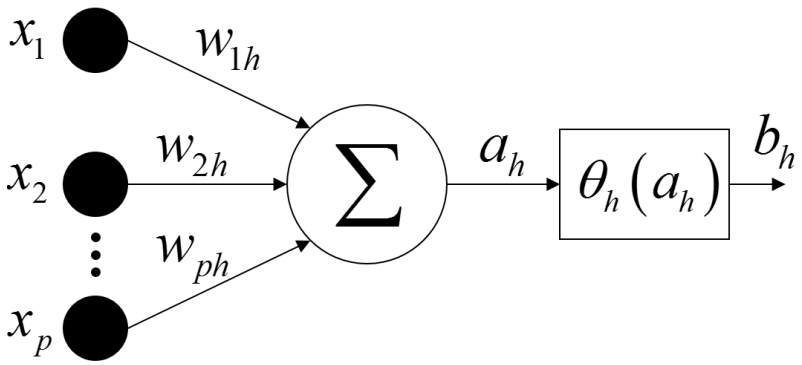
- Model mimicking the information processing of biological brains.
- Network of small processing units (**neurons**) joined to each other by weighted connections (**synapses**).



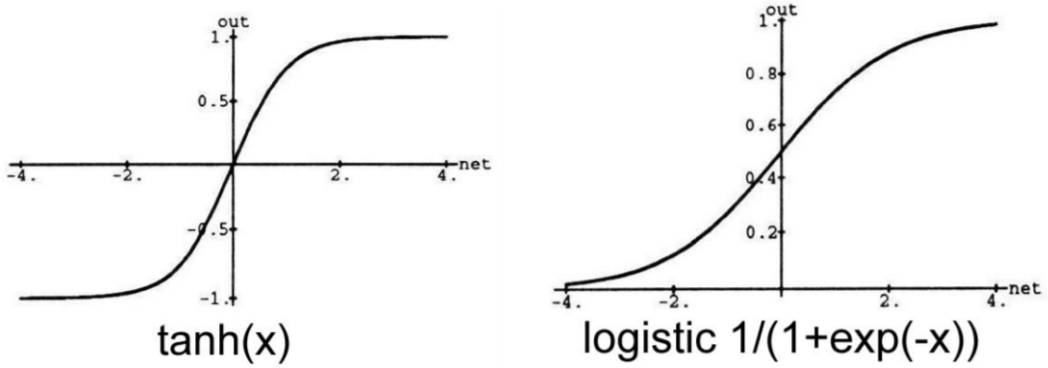
Deep learning for probabilistic forecasting

Artificial Neural Network

Neuron model



Activation function θ_h



Objective: Find optimal weights → Backprop

Deep learning for probabilistic forecasting

Backpropagation

Algorithm for training Neural Networks

1. Forward Pass: Given inputs and current weights, compute outputs of the ANN
2. Loss Function: Outputs is compared with actual observations using a pre-defined error function (e.g. SSE)
3. Backward Pass: Error values are propagated backwards
4. Weights Update: Gradient Descent algorithm

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. In Neurocomputing: Foundations of research, pages 673–695. MIT Press, Cambridge, MA, USA, 1988.

Deep learning for probabilistic forecasting

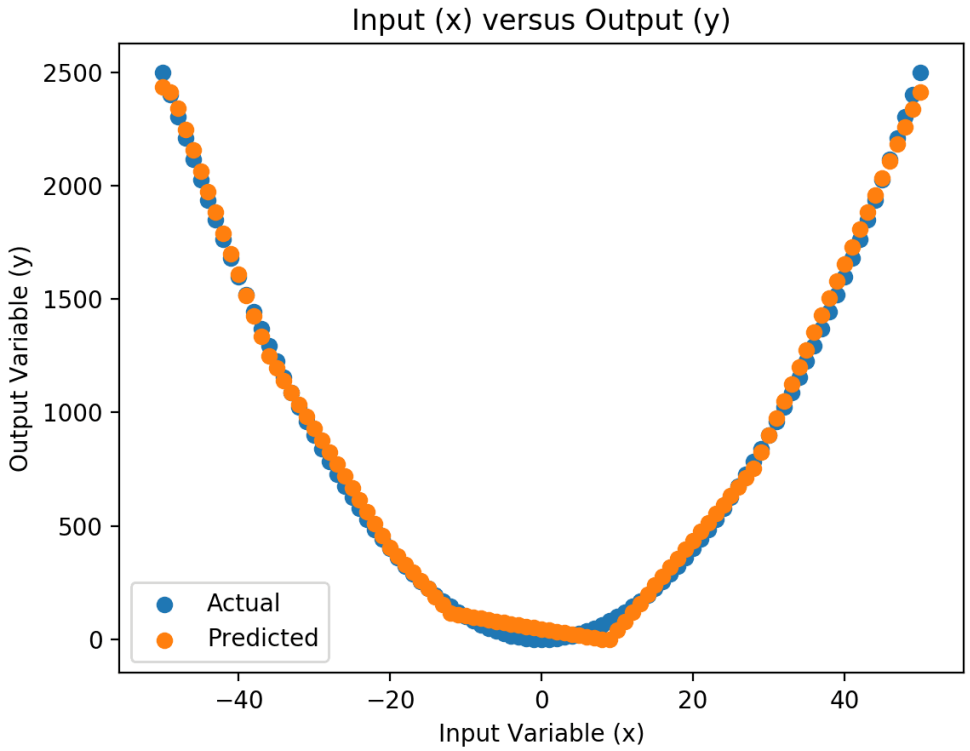
Towards Modeling Nonlinearities

Any combination of linear operators is itself a linear operator.

This contrasts with nonlinear networks, which are able to gain considerable power by building up progressively higher level representations of data using **deep architecture**.

➔ANN are theoretically able to learn any nonlinear relationship between a dependent variable and its predictors.

➔Need of many historical data !!



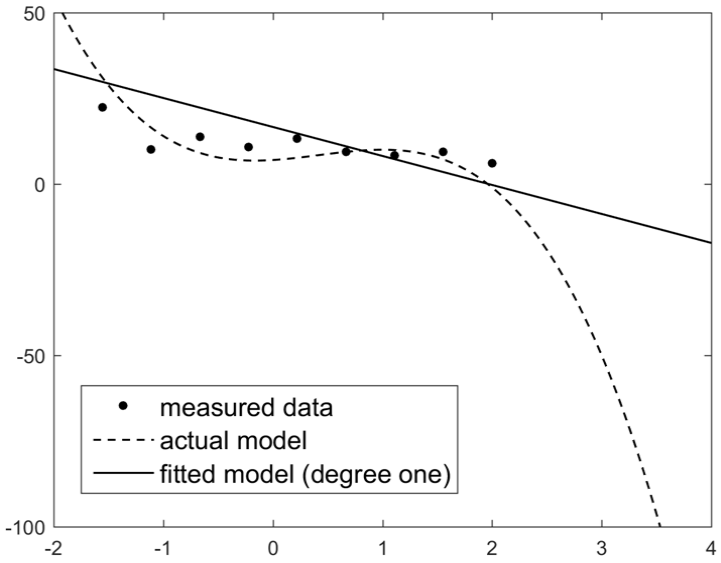
CHAPTER 4 A visual proof that neural nets can compute any function http://neuralnetworksanddeeplearning.com/chap4.html

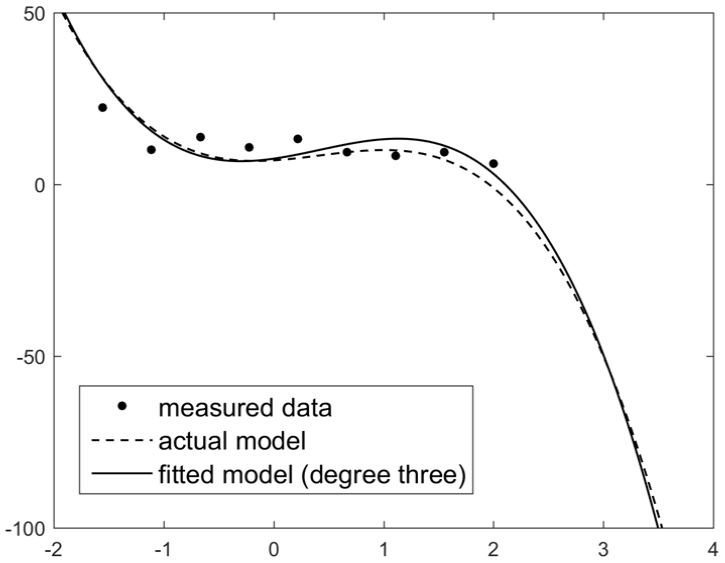
Deep learning for probabilistic forecasting

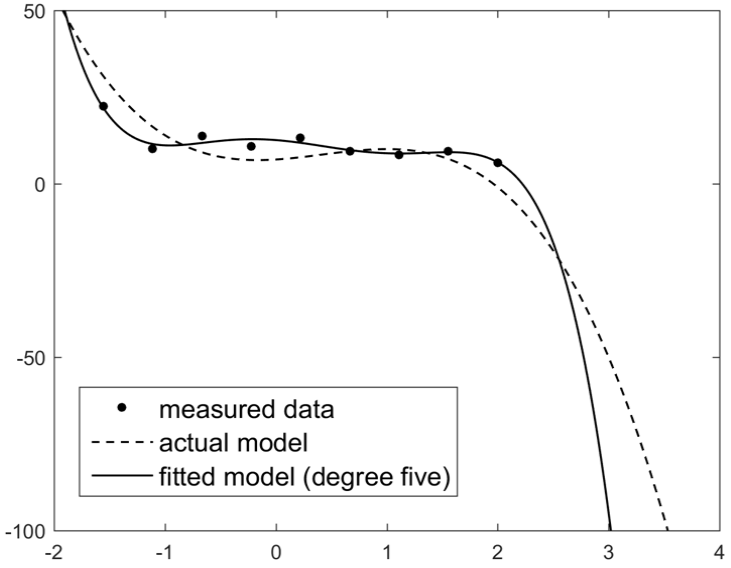
How to avoid overfitting?

Overfitting is a modelling error that arises when the model is too closely adapted to a limited set of data points.

Origins: small training dataset or improper model complexity







Deep learning for probabilistic forecasting

Regularization techniques

1) Early stopping: divide data into training and validation sets
2) Input noise
3) Weight noise

Noise reduces the precision with which the weights must be described.

Simpler networks are preferable because they tend to generalize better.

Deep learning for probabilistic forecasting

Features Selection

1) Correlation study – Keep inputs with a sufficient degree of dependence on the predicted variable
2) Embedded into optimization procedure – Select input combination that minimizes error (Feature importance in RF and GBRT)
3) Hybridization of both
4) “Random Bar” based method

Deep learning for probabilistic forecasting

Damien Benveniste, PhD • 2nd
Machine Learning Tech Lead at Meta (Facebook)
2d •

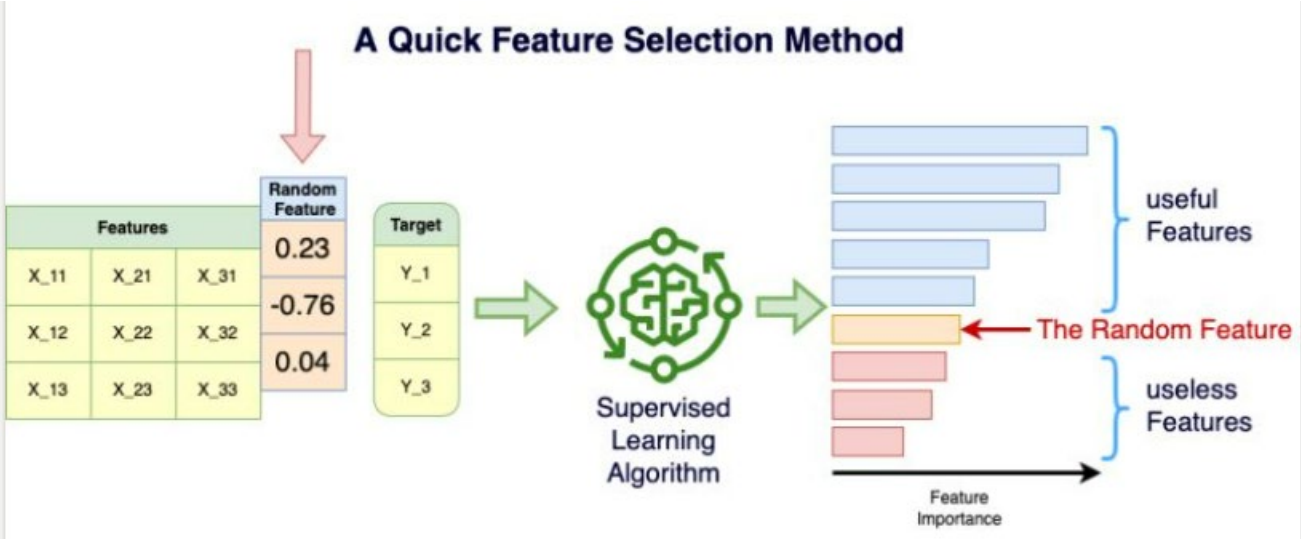
+ Follow

This is a technique I like to perform a quick FEATURE SELECTION for Machine Learning applications. I tend to call it the "Random Bar" method! Let's assume you have a feature set X and a target Y. Let's create a random vector V (for example np.random.normal(size=(1, 100))) and append that vector as a new feature to X:

X' = [X, V]

X' is just the original feature set with additionally the new random feature. Keep in mind that this new feature cannot possibly help to predict the target Y since it is random! Now, take that data (X', Y) and train a Supervised Learning algorithm with a Feature Importance measure that is relevant for you application. Intuitively, the mean entropy gain per split of tree based algorithms (Random Forest, Xgboost, ...) is a convincing measure of feature importance to me. The statistical fluctuation of the data is such that even the random feature will be attributed a non-zero feature importance by the algorithm, but we know it is artificial. Any feature with a lower feature importance than the random feature has to be useless to predict the target and the features with a higher feature importance are at least better than random noise at predicting the target.

This is especially useful if you have thousands of features and you want to weed out quickly the ones that won't have any impact in the learning process. This is also a method that can be used for highly non-linear data as opposed to LASSO (for example) that tends to only understand linear relationships in the data. The random feature is a "Random Bar" because this is the minimum bar a feature needs to beat to be a part of the potentially useful features set. Now it doesn't mean there are not additional features that could be beneficial to further remove to optimize your model. Do you know if this method has a more jargon-y name?



Deep learning for probabilistic forecasting

Architecture Optimization

1) Architecture: number of hidden layers, of neurons within each layer, etc.
2) Regularization technique and its parameters
3) Optimization algorithm and its parameters

Methodology:

Grid search, random search, sequential learning

Bergstra, J., Bardenet, R., Bengio, Y., and Kegl, B. **Algorithms for hyper-parameter optimization**. In NIPS'2011.

Deep learning for probabilistic forecasting

How to Incorporate Context?

- In the context of time series prediction, temporal dependencies have to be adequately taken into account
- Traditionnal ANN intrinsically treats task as spatial

➔Architecture has to be tailored !

Time Delay Neural Network, NARX model, Recurrent Neural Network, Long Short Term Memory, Transformer……

Deep learning for probabilistic forecasting

Time Delay Neural Network

The output in time t is based on the inputs in times (t-1),(t-2), ..., (t-n).

$$y(t) = \mathrm{f}\left(u(t), u(t-1), \ldots, u(t-n_u)\right)$$

Optimal **window size is task-dependent** and requires finding trade-off between sufficient temporal information while avoiding overfitting

Deep learning for probabilistic forecasting

NARX model

The output in time t is based on the inputs **& outputs** in times (t-1),(t-2), ..., (t-n)

$$y(t) = \mathrm{f}\left(y(t-1),\ldots,y(t-n_y),u(t),u(t-1),\ldots,u(t-n_u)\right)$$

Optimal **window size is task-dependent** and requires finding trade-off between sufficient temporal information while avoiding overfitting

Recurrent Neural Network

With purely feedforward networks, some of the temporal structure in time series is ignored.

In recurrent neural networks (RNN) there may be connections between the neurons in a layer

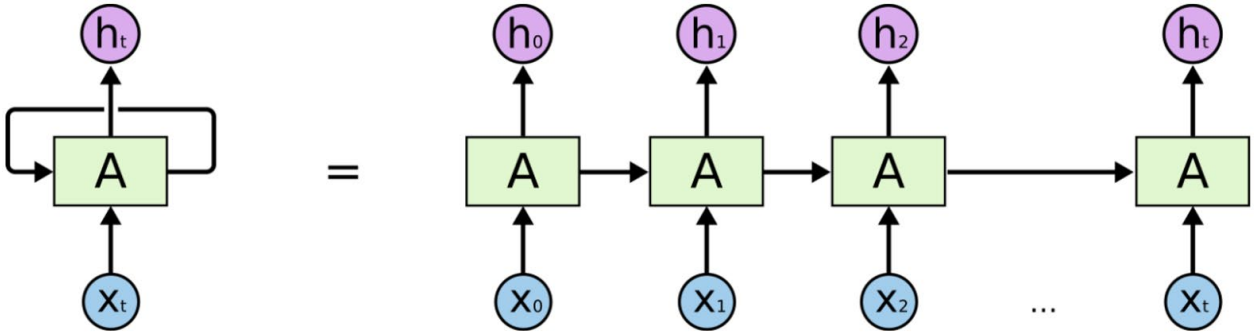
The principle is to preserve the sequential information through the dynamic induced in the network's hidden state

Deep learning for probabilistic forecasting

Recurrent Neural Network

RNN = Neural Networks with feedback loops.

RNN are harder to train due to higher complexity → Backpropagation Through Time



Vanishing Gradient Problem: back-propagated errors during training either fades or blows up over time depending on the neurons activation function.

Paul J. Werbos. Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10):1550–1560, 1990

Deep learning for probabilistic forecasting

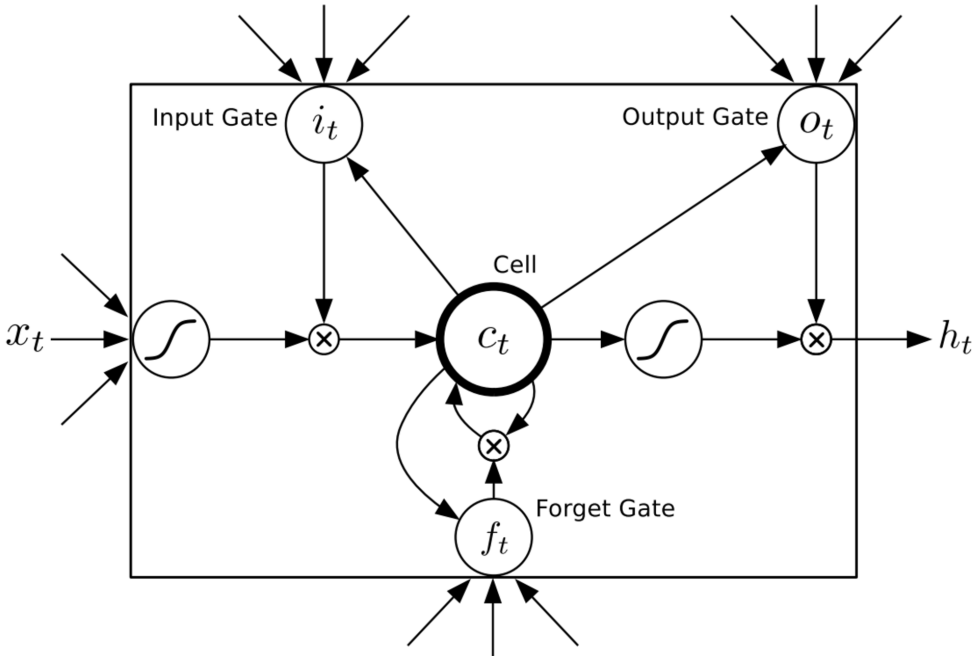
Long Short Term Memory

The LSTM network controls the flow of information through the hidden layer using gating units

Input Gate = write

Forget Gate = reset

Output Gate = read



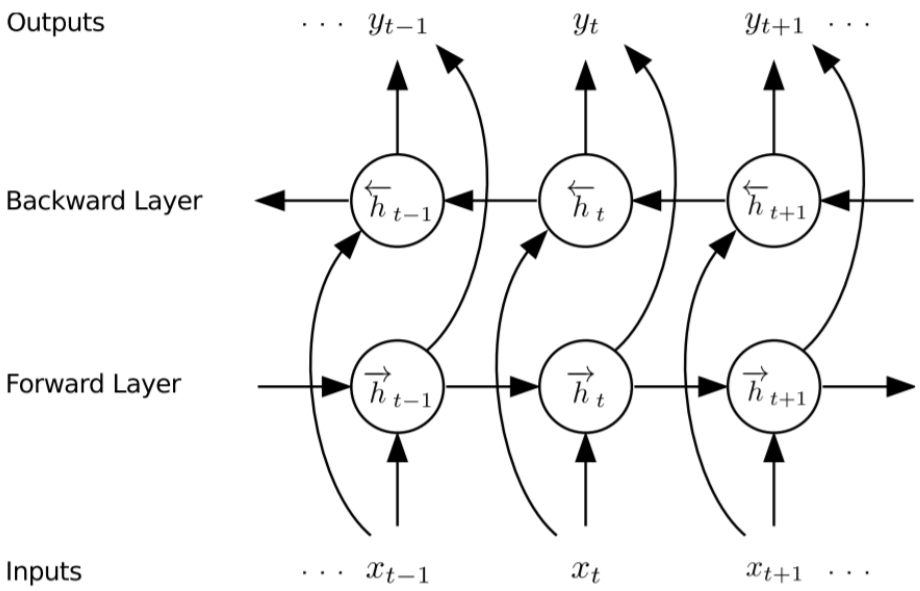
RNN process inputs in temporal order and ignore information contained in future context (causality violation?)

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8):1735–1780, 1997.

Deep learning for probabilistic forecasting

Bidirectional RNN

Bidirectional RNNs exploit past and future context by processing the data in both directions with two separate hidden layers.

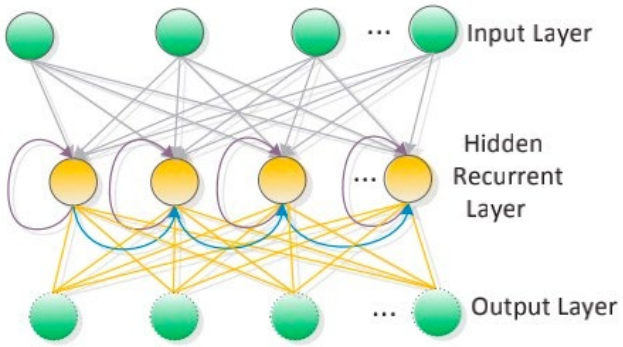


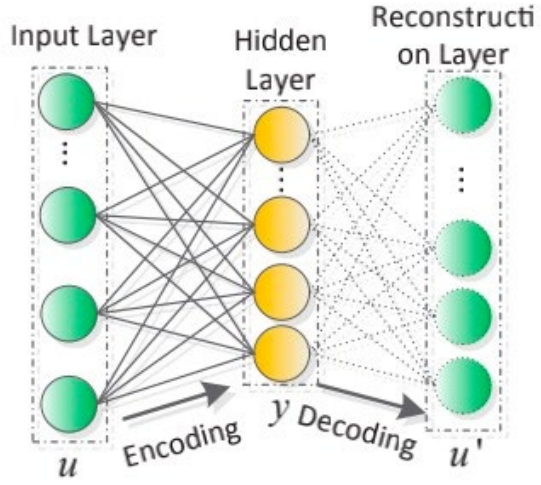
Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45:2673–2681.

Deep learning for probabilistic forecasting

Other DNN models:

- Transformer
- Convolutional neural networks
- Auto-encoder
- Auto Machine Learning?
- ……



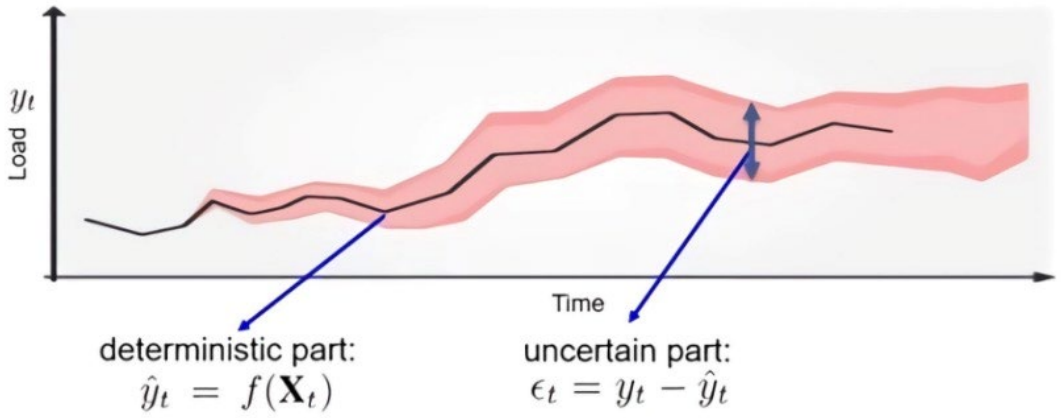


Deep learning for probabilistic forecasting

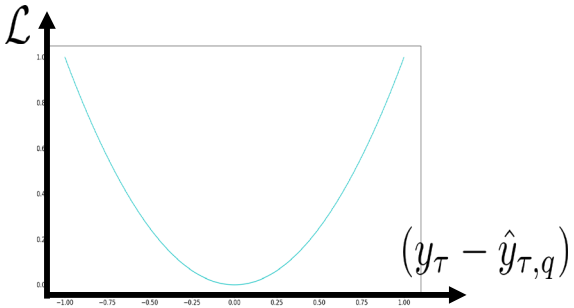
Probabilistic forecasting methods

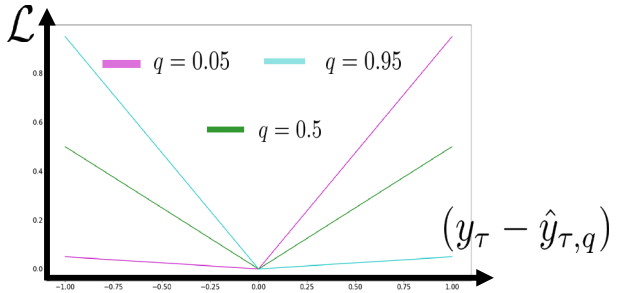
Probabilistic forecasts can be in the form of quantiles, intervals, or density function.

Category	Methods
Parametric method	Auto-regression model
	Maximum likelihood
	Bayesian approach
Nonparametric method	Quantile regression
	Bootstrapping method
	Lower upper bound estimate
	Gradient boosting
	Kernel density estimation
	Analog ensemble



- **Deterministic regression**
- **Quantile regression**



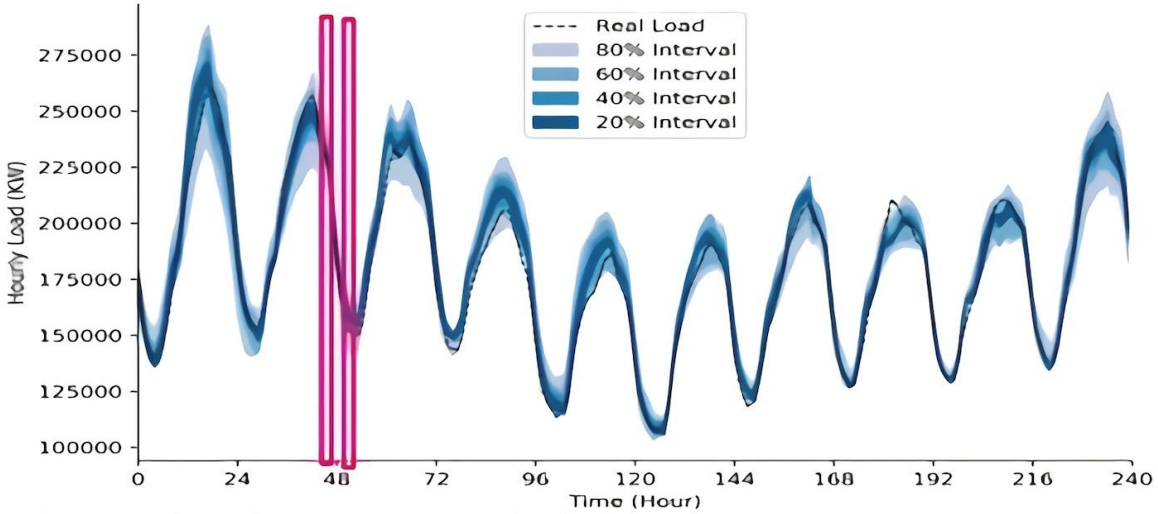


- **Supervised learning:** $\Theta^* = \underset{\Theta}{\operatorname{argmin}} \sum_{i \in \mathcal{I}} \mathcal{L}(\mathcal{Y}_i, f_\Theta(\mathcal{X}_i))$

model parameters

Deep learning for probabilistic forecasting

Generative neural networks



Traditional PLF can only capture the probability distribution of the load individually in each period and can not integrate dependencies among different periods.

Dependencies across periods

Generating scenarios is an effective way to capture such dependencies!

Renewable energy scenarious generation?

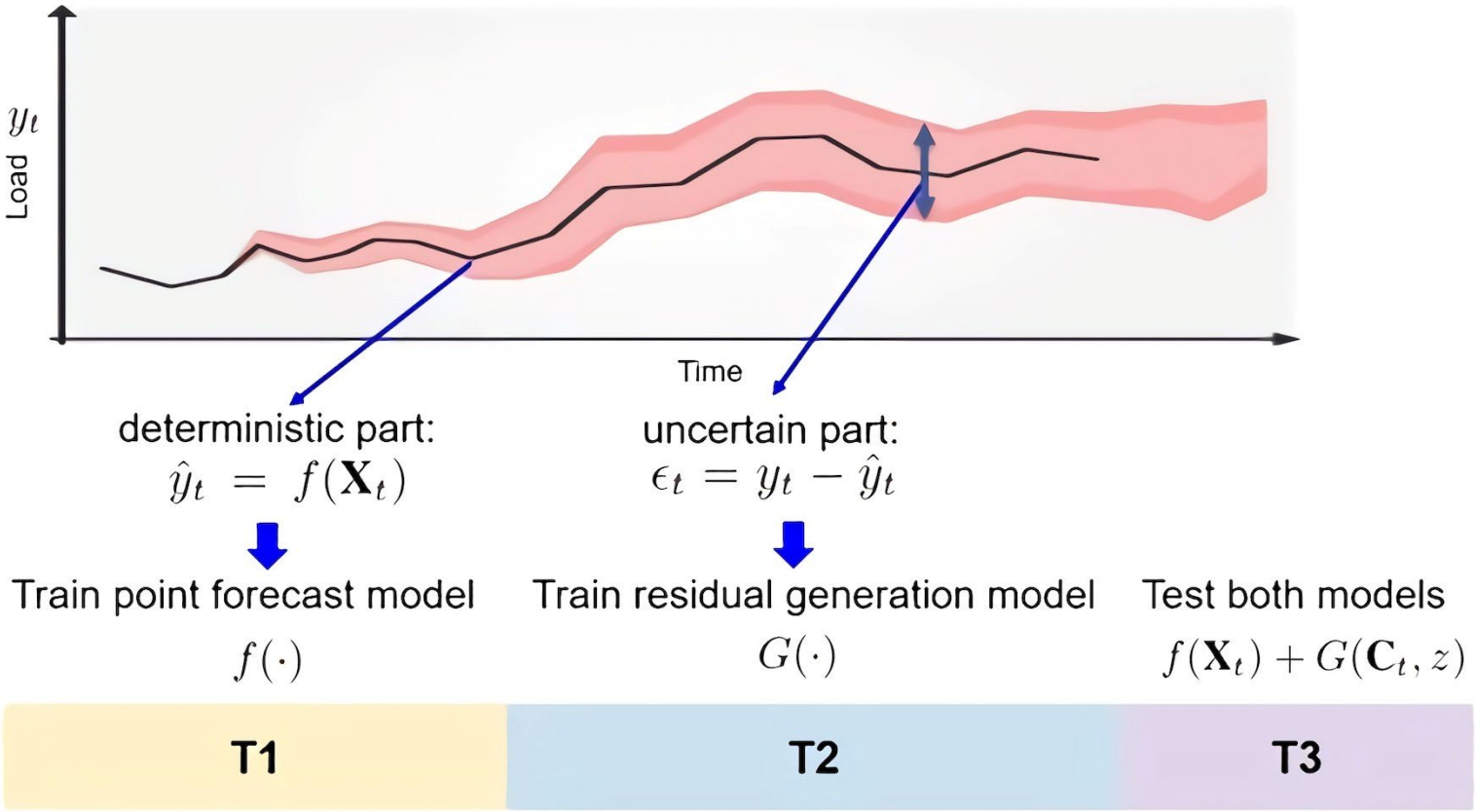
Long-term uncertainties?

Our work focuses on the short-term load scecarios generation

Deep learning for probabilistic forecasting

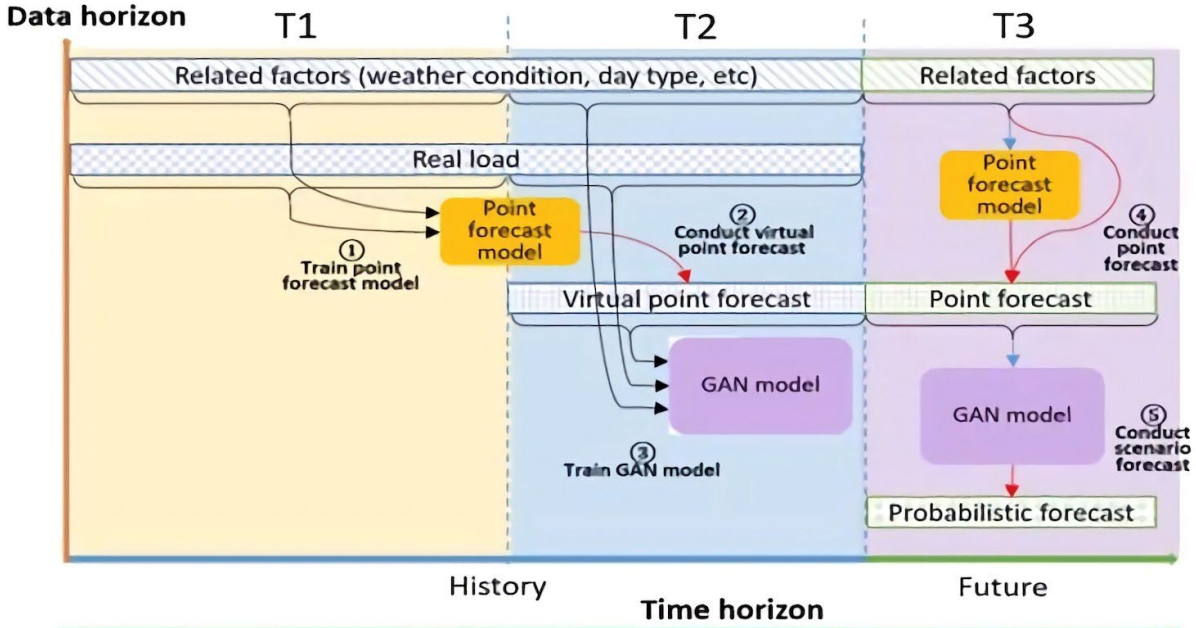
Generative neural networks

From the perspective of forecasting, the electrical load contains two parts:



Deep learning for probabilistic forecasting

Generative neural networks

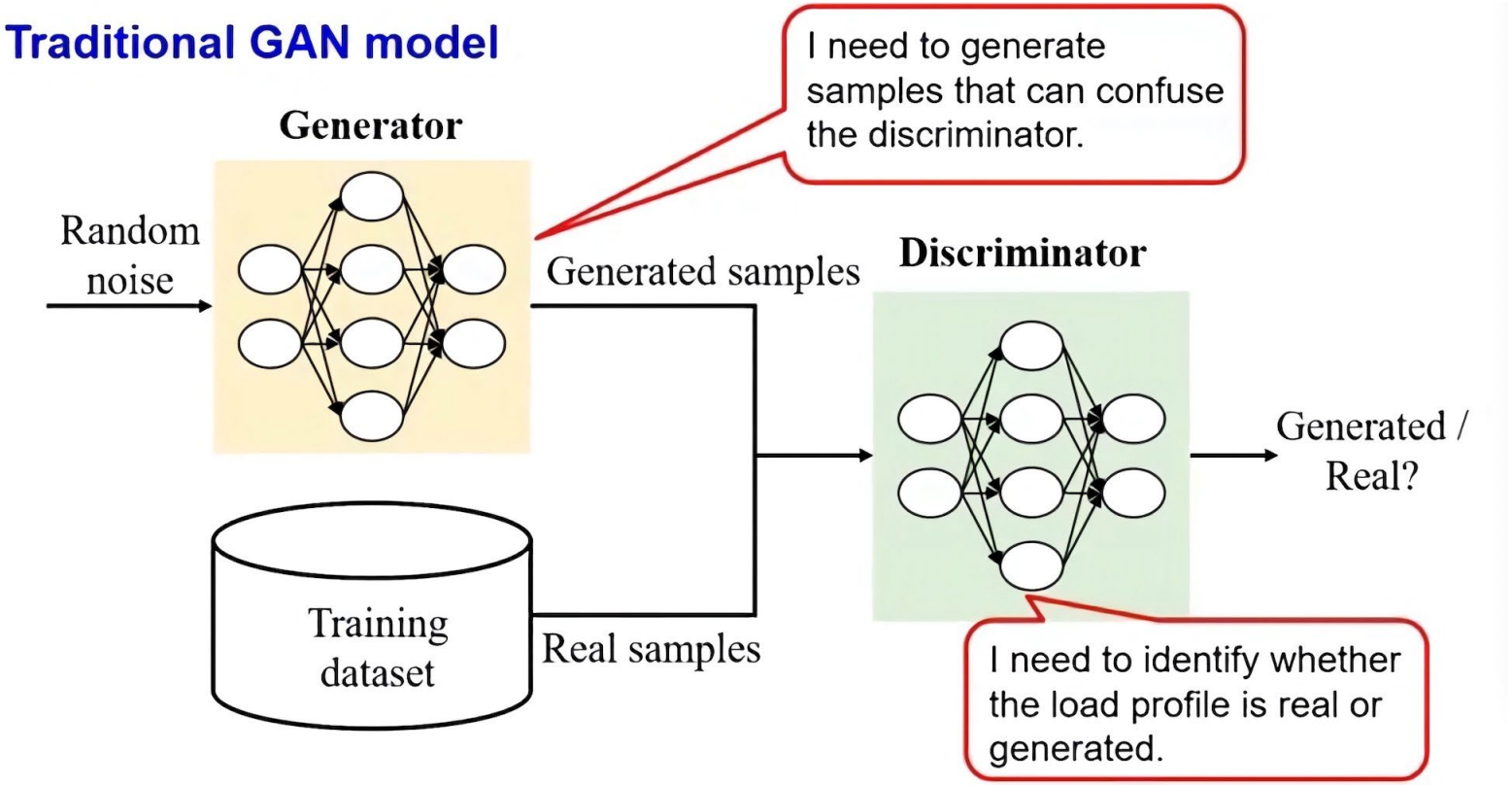


➢ **Train** $f(\mathbf{X}_t)$ using point forecasting models such as MLP/SVR/RF/GBRT	➢ **Test** $f(\mathbf{X}_t)$ ➢ **Calculate** ϵ_t ➢ **Train** $G(\mathbf{C}_t, z)$	➢ **Test** $f(\mathbf{X}_t)$ ➢ **Test** $G(\mathbf{C}_t, z)$ ➢ **Final forecasts** $f(\mathbf{X}_t) + G(\mathbf{C}_t, z)$

Formulate the generation model $G(\cdot)$ using GAN !

Deep learning for probabilistic forecasting

Generative neural networks



The **adversarial game** between these two neural networks can be presented as a min-max optimization model:

$$\max_{\theta_D} \mathbb{E}_{S_r}[\log(D(s_r;\theta_D))] + \mathbb{E}_Z[\log(1-D(G(z;\theta_G);\theta_D))]$$

Deep learning for probabilistic forecasting

Generative neural networks

Evaluation Criteria

From uncertainty perspective: **Pinball loss (PL)** and **Winkler Score (WS)** assess the calibration and sharpness simultaneously.

$$\mathrm{PL}\left(\hat{y}_{t,q}, y_t\right) = \begin{cases} \left(y_t - \hat{y}_{t,q}\right)q & \hat{y}_{t,q} \le y_t \\ \left(\hat{y}_{t,q} - y_t\right)(1-q) & \hat{y}_{t,q} > y_t \end{cases}$$

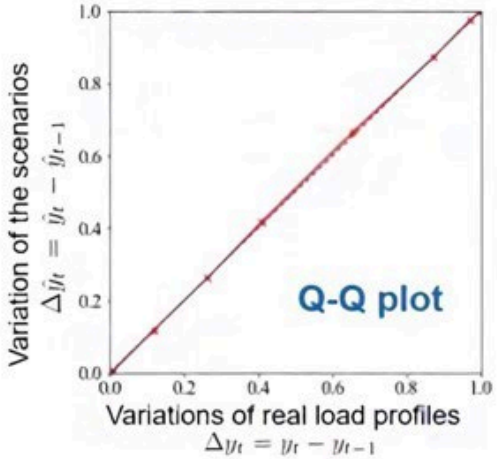
Performance of overall quantiles

$$\mathrm{WS}(L_t, U_t, y_t) = \begin{cases} \delta_t + 2(L_t - y_t)/\alpha & y_t \le L_t \\ \delta_t & L_t < y_t < U_t \\ \delta_t + 2(y_t - U_t)/\alpha & U_t \le y_t \end{cases}$$

Performance of extreme quantiles

From variation perspective:

the **Q-Q plot** visually evaluate the similarity of the distributions of the variations of the real load profiles and the generated scenarios.



Deep learning for probabilistic forecasting

Generative neural networks

Case Studies: Performance w.r.t Uncertainty

Point Forecasts	Uncertainty Modeling	PL	WS(a=0.2)	WS(a=0.1)
AVE	Proposed	**12.38**	**180.1**	262.14
	CWGAN	13.7	191.49	261.59
	QRF	14.23	189.64	**231.84**
	QGBRT	14.05	190.66	243.66
SVR	Proposed	**12.55**	**182.04**	259.84
	CWGAN	12.78	190.91	281.9
	QRF	14.5	194.66	**240.76**
	QGBRT	14.75	201.88	255.55
RF	Proposed	**12.91**	**183.32**	260.07
	CWGAN	13.1	187.58	263.7
	QRF	14.44	194.06	242.99
	QGBRT	13.94	186.66	**233.99**
GBRT	Proposed	**12.24**	**172.15**	236.34
	CWGAN	13.11	180.22	235.96
	QRF	14.06	183.18	**223.12**
	QGBRT	14.36	189.39	236.59

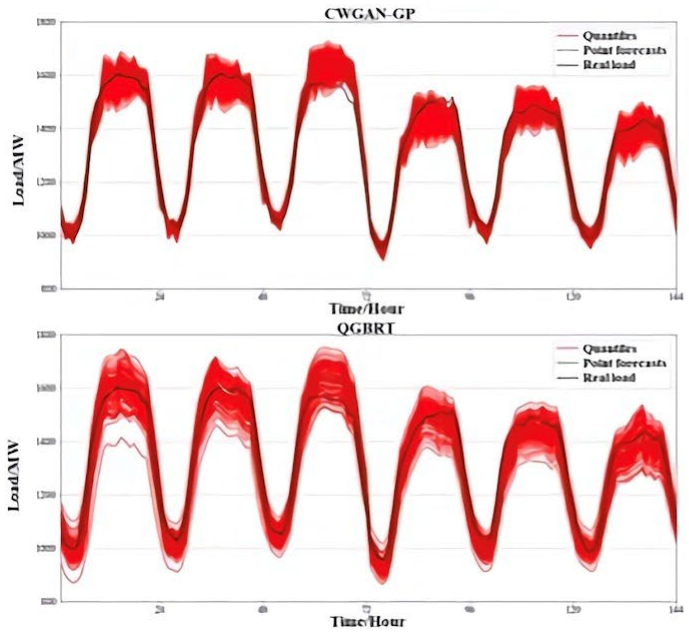
Comparison among different PLF methods

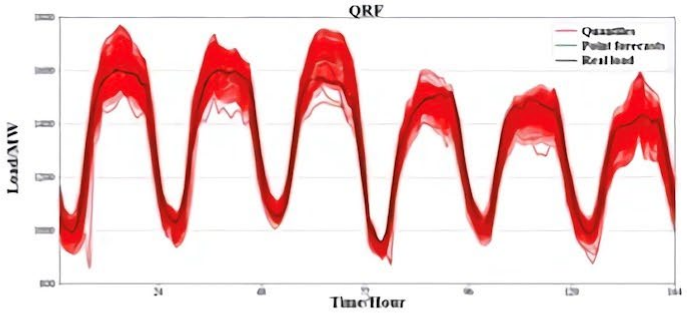
- For different point forecasts, our proposed CWGAN-GP model outperforms the CWGAN model, QRF, and QGBRT in terms of **PL and WS (a= 0.2)**.
- However, QRF instead of the CWGAN-GP model performs better in terms of **WS (a= 0.1)**.

Deep learning for probabilistic forecasting

Generative neural networks

Case Studies: Investigation on Extreme Quantiles





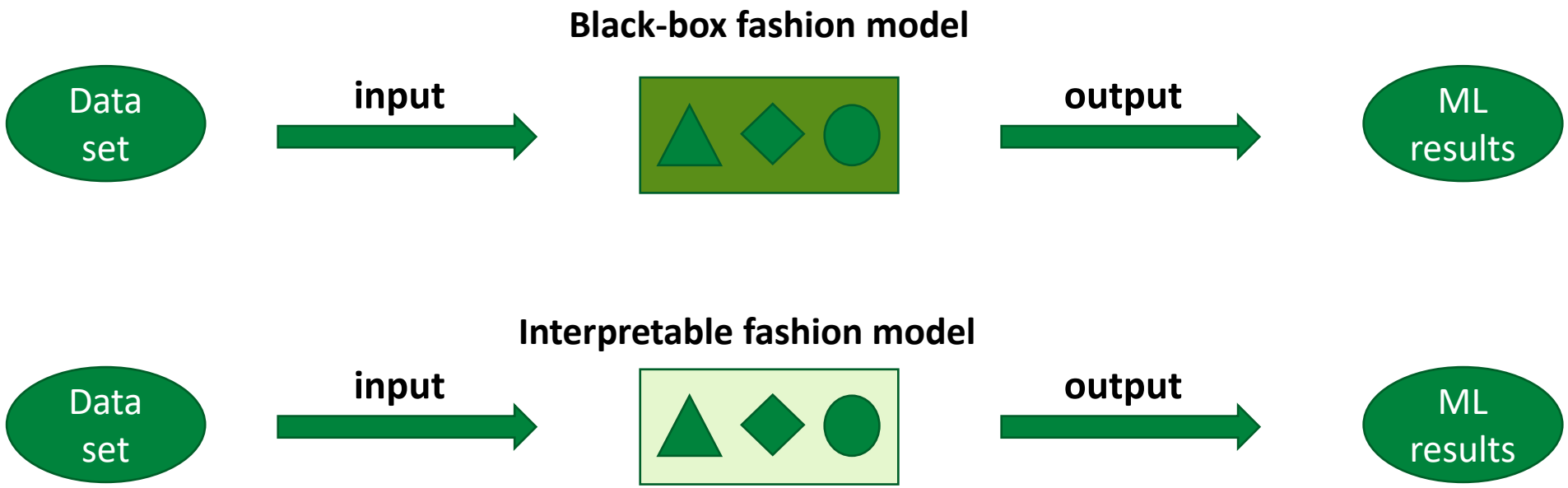
Quantile forecasts obtained by CWGAN-GP, QGBRT, and QRF

- The scenarios generated by CWGAN-GP model are more concentrated compared with those by QGBRT and QRF.
- Such results suggest that it is hard for the CWGAN-GP model to generate extreme scenarios

Deep learning for probabilistic forecasting

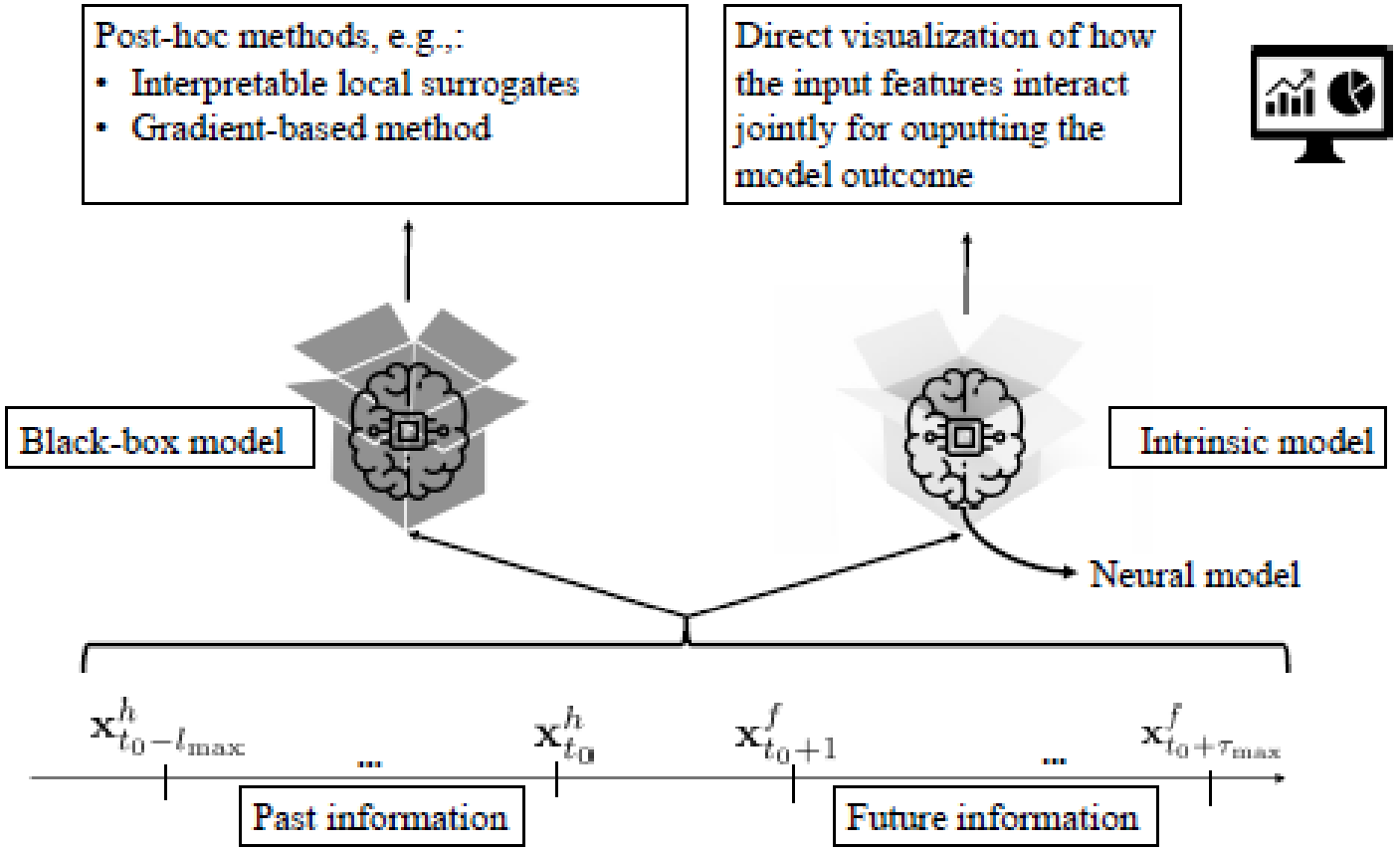
Explainable machine learning

Providing insights into inner decision-making processes



Deep learning for probabilistic forecasting

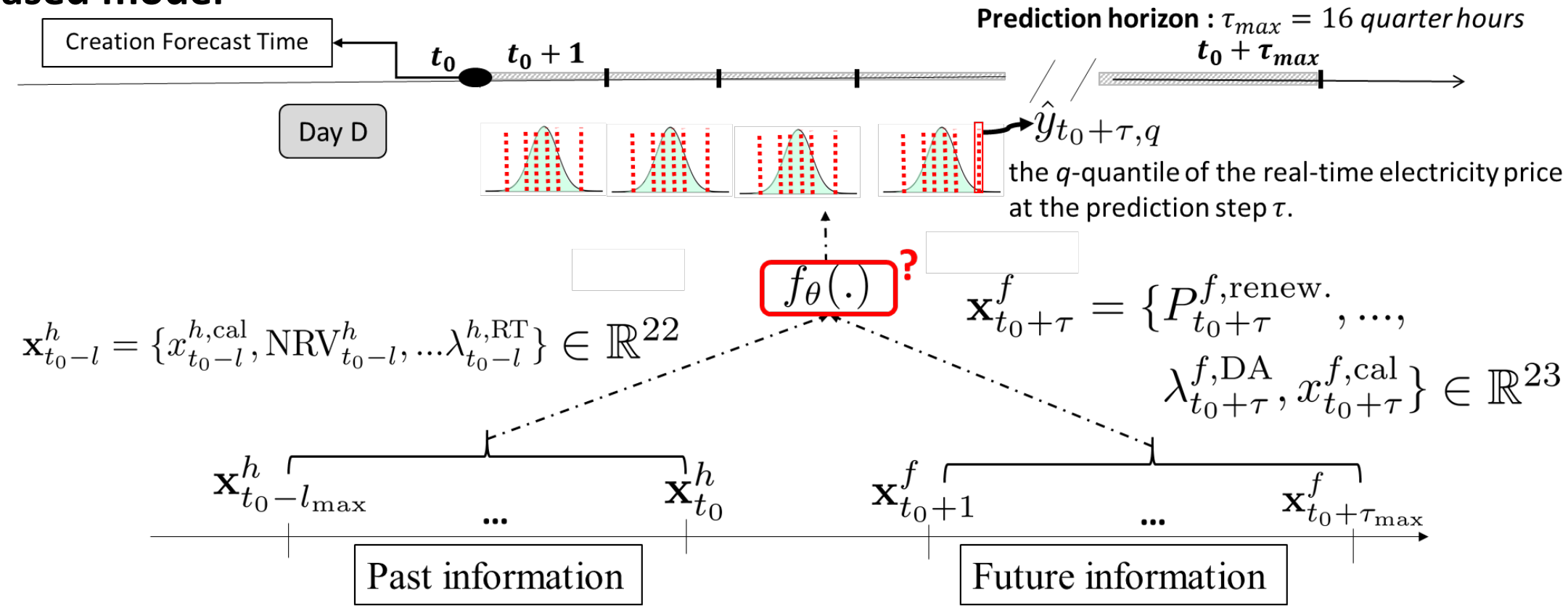
Explainable machine learning



Deep learning for probabilistic forecasting

Explainable machine learning

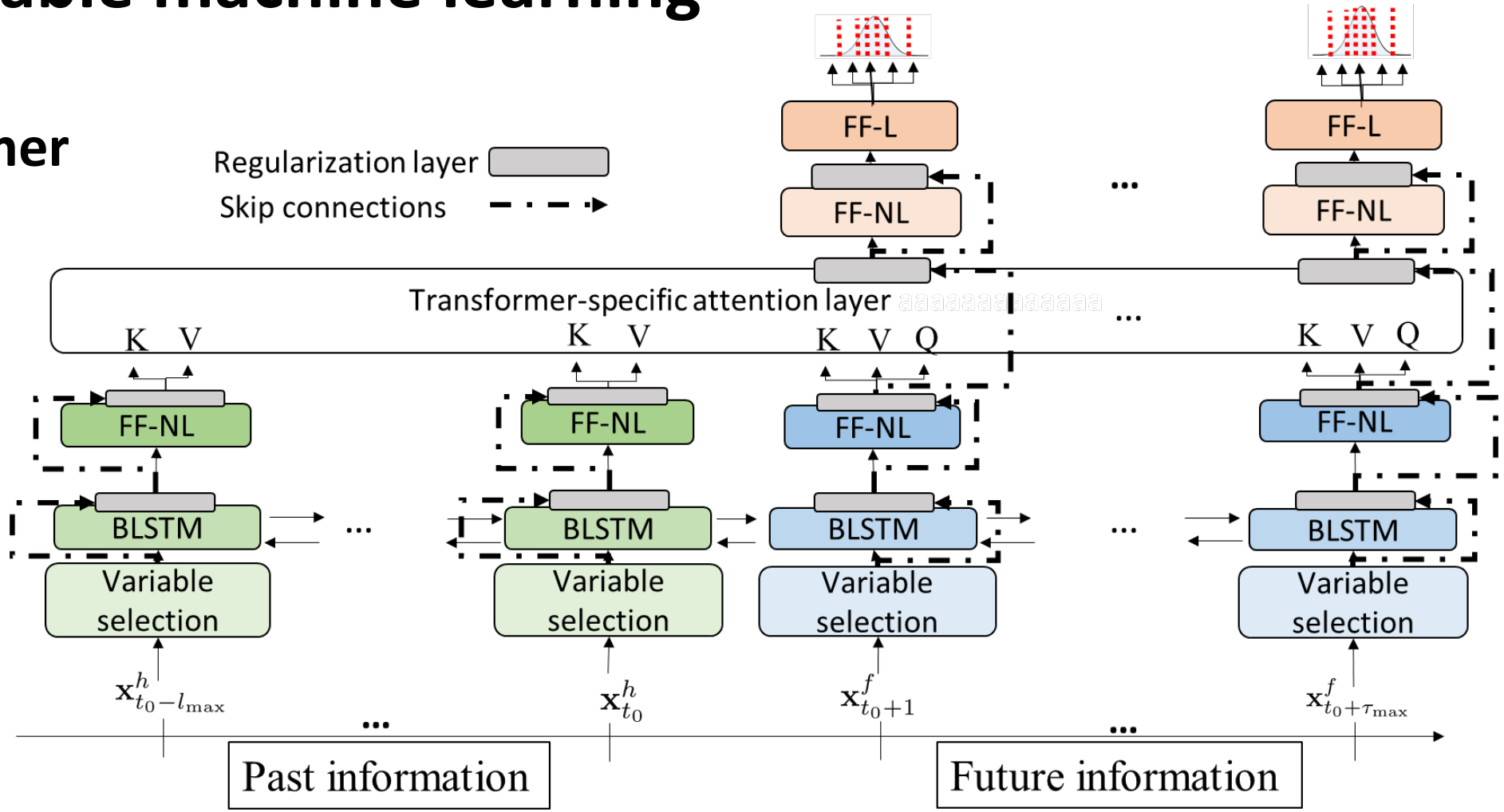
Interpretable probabilistic forecasting of real-time electricity prices using a Transformer-based model



Deep learning for probabilistic forecasting

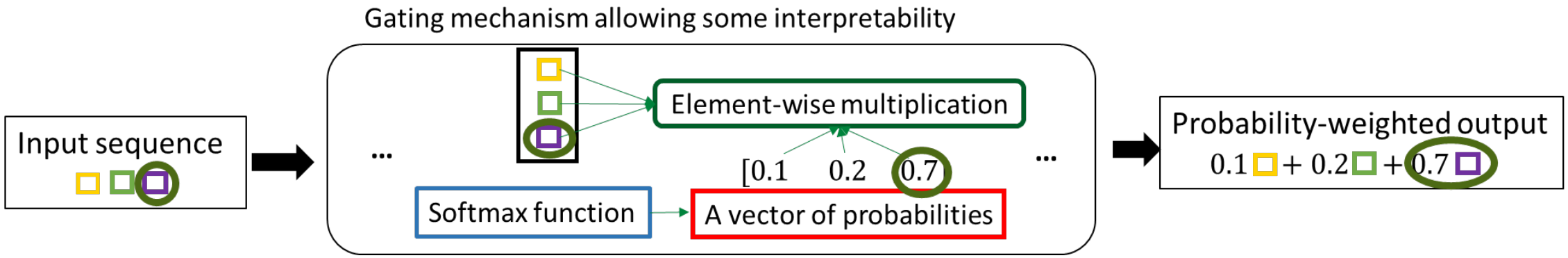
Explainable machine learning

Transformer

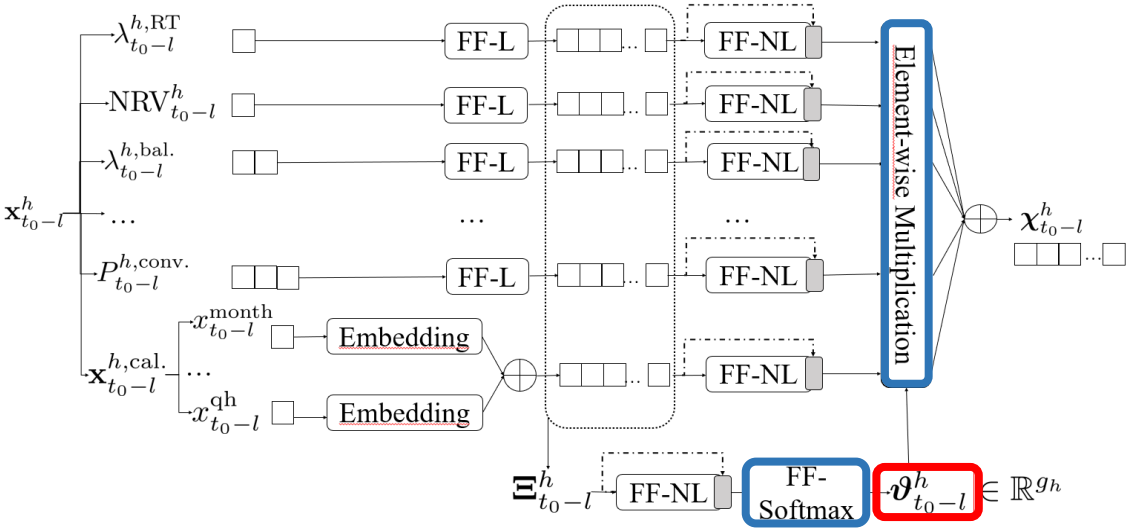


Deep learning for probabilistic forecasting

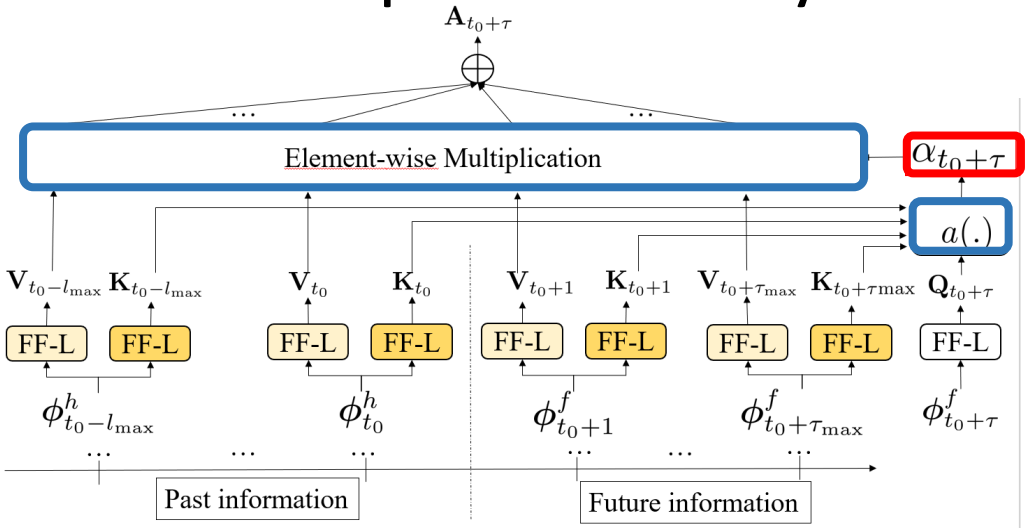
Explainable machine learning



1. Variable selection



2. Transformer-specific attention layer



Deep learning for probabilistic forecasting

Explainable machine learning

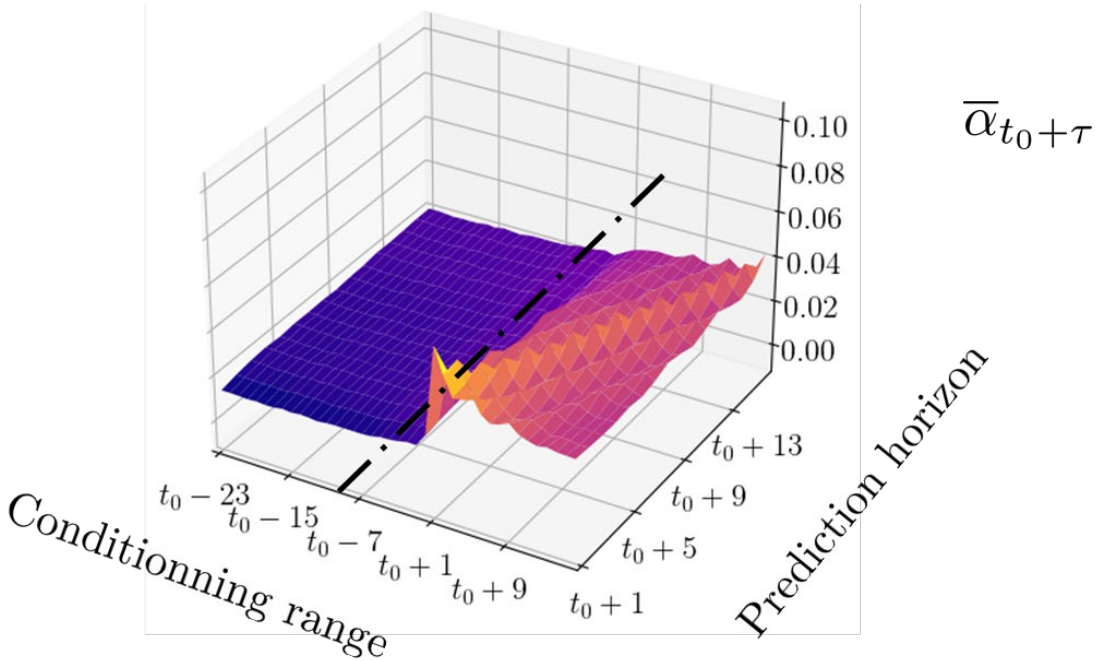
Global temporal input feature importance

Test set: the year 2019, which is approximatively 35,000 novel (unseen) input conditions.

- **Averaged input importance**

	$\overline{\vartheta}^h$	$\overline{\vartheta}^f$
λ^{RT}	**0.27**	/
NRV	**0.26**	/
$\lambda^{\mathrm{bal.}}$	0.05	/
ϕ	0.03	/
$L^{\{h,f\}}$	0.01	0.03
$P^{\{h,f\},\mathrm{renew.}}$	**0.11**	0.04
$P^{\{h,f\},\mathrm{conv.}}$	0.05	< 0.01
$x^{\mathrm{cal.}}$	**0.22**	0.02
λ^{DA}	/	**0.86**
$\lambda^{\mathrm{f,bal.}}$	/	0.05

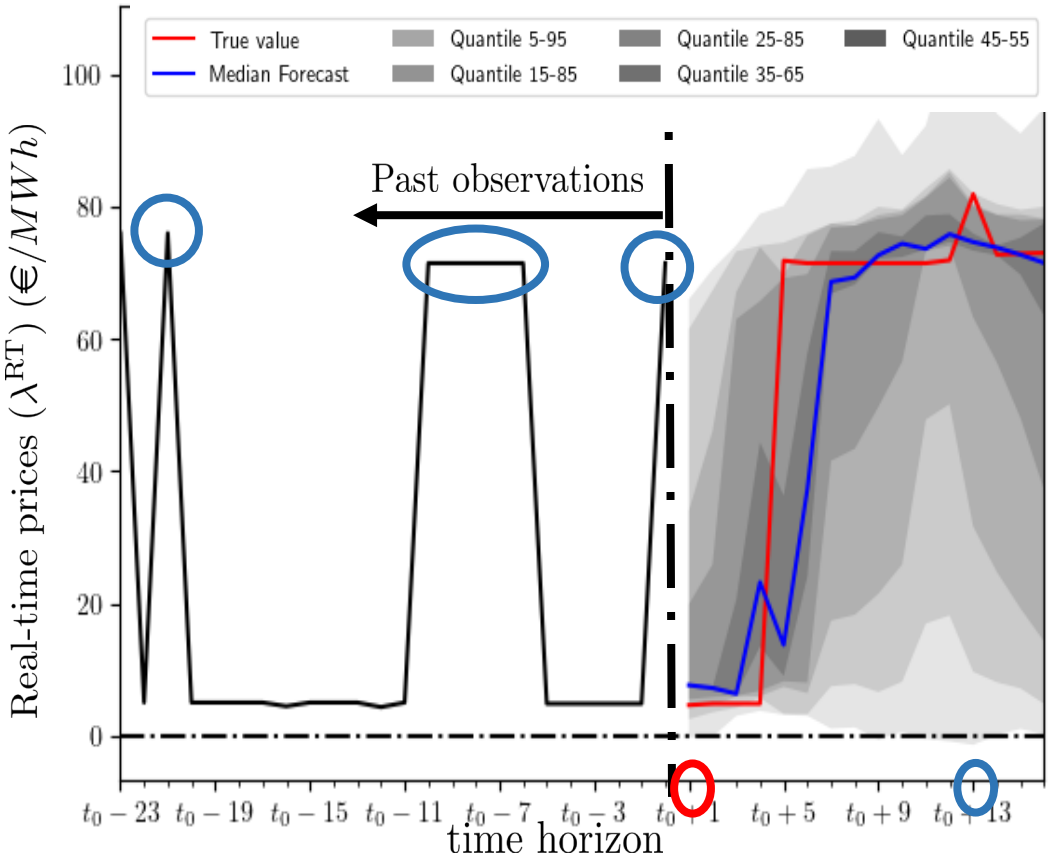
- **Averaged temporal attention**

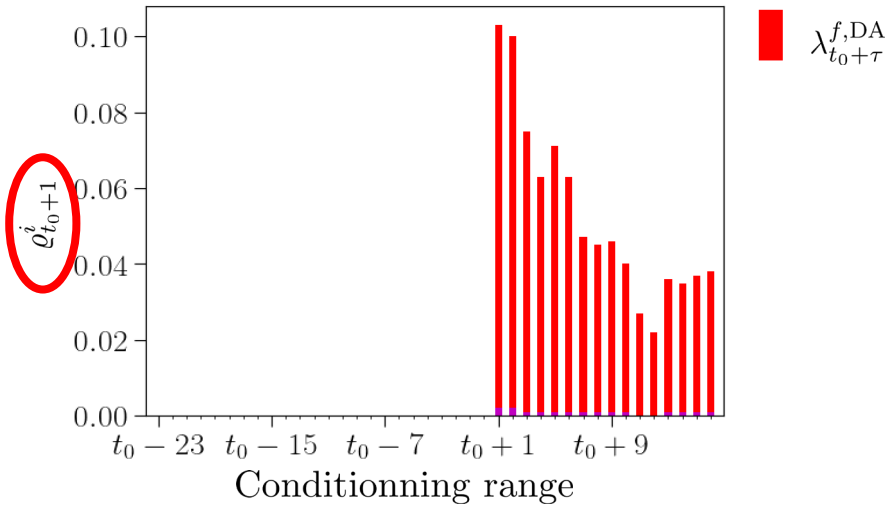


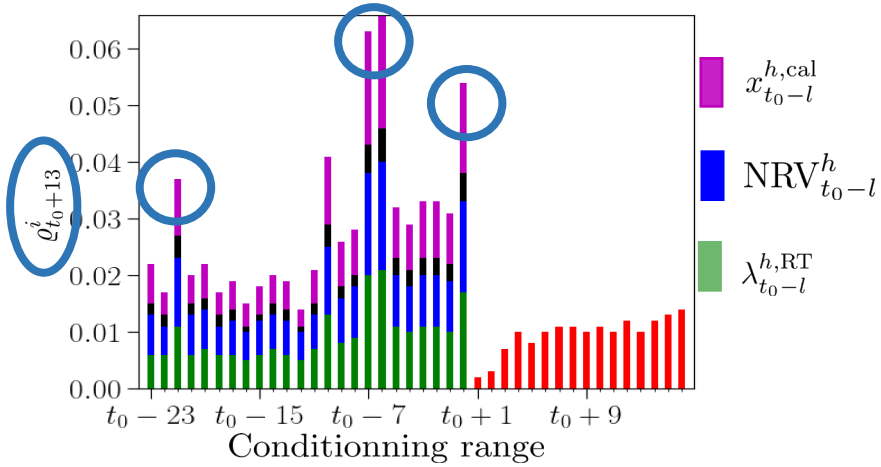
Deep learning for probabilistic forecasting

Explainable machine learning

Case-specific temporal input feature importance







Deep learning for probabilistic forecasting

Explainable machine learning

Enriching the neural network architecture with interpretable layers does not hinder the prediction performance, while being beneficial for:

- the designer, who can check that the model does not exploit artifacts in the data,
- the end-user, who will be better equipped for making a decision

Some perspectives:

- Extensive benchmark analysis investigating both post-hoc and inherent interpretable methods.

- Developing natural language interpretations for leveraging intelligibility for non-expert users.

Deep learning for probabilistic forecasting

Summary:

- Pro and cons of different networks;
- Physical knowledge embedding;
- Explainable/Interpreterable networks;
- Do we really need DEEP networks???

With time-covariates				
Dataset		DeepGlo	GBRT(Naive)	GBRT(W-b)
Electricity	RMSE	141.285	175.402	**119.051**
	WAPE	0.094	0.288	**0.089**
	MAE	53.036	143.463	**50.150**
Traffic	RMSE	0.026	0.038	**0.014**
	WAPE	0.239	0.495	**0.112**
	MAE	0.013	0.027	**0.006**
PeMSD7	RMSE	6.490*	8.238	**5.194**
	WAPE	0.070	0.100	**0.048**
	MAE	3.530*	5.714	**2.811**
Exchange-Rate	RMSE	0.038	0.079	**0.016**
	WAPE	0.038	0.450	**0.013**
	MAE	0.029	0.066	**0.010**

(*) Results reported from the original paper.

Elsayed S, Thyssens D, Rashed A, et al. Do we really need deep learning models for time series forecasting?[J]. arXiv preprint arXiv:2101.02118, 2021.

Combining probabilistic forecasts

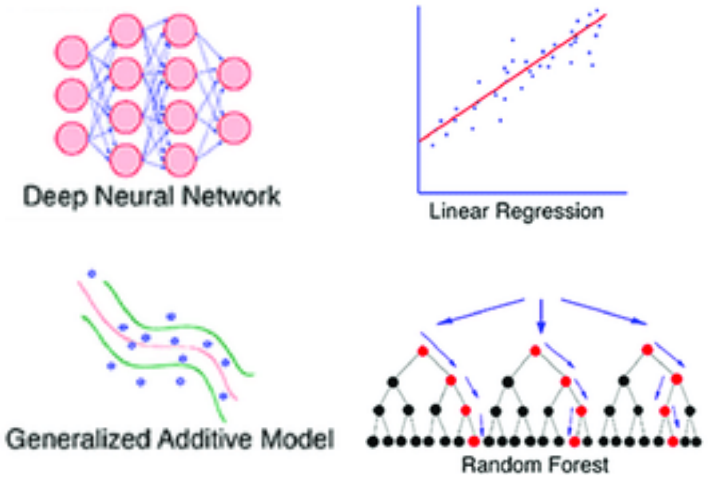
Combining probabilistic forecasts

Ensemble Learning

*In statistics and machine learning, **ensemble methods** use **multiple learning algorithms** to obtain **better predictive performance** than could be obtained from any of the constituent learning algorithms alone [1].*

Western Phrase: Two heads are better than one;

Chinese Saying: Three vice-generals are equal to one Zhuge Liang



- Which method is the best?
- Is it possible to combine these methods?

[1] https://en.wikipedia.org/wiki/Ensemble_learning

Combining probabilistic forecasts

Problem Formulation

Loss function: MAPE, RMSE, etc.

Combined forecast

Determine the weights

$$\min_{\omega_n} \quad TL = \sum_{t=1}^{T} L\left(\sum_{n=1}^{N} \omega_n F_{n,t}, y_t\right)$$

Real load

$$\text{s.t.} \quad \sum_{n=1}^{N} \omega_n = 1, \quad \omega_n \geq 0$$

Summation and non-negative constraints

Combining probabilistic forecasts

Ensemble Learning for probabilistic forecasting

Various ensemble methods have been studied to combine multiple point forecasts.

However, **combining probabilistic load forecasts is a rarely touched area**.

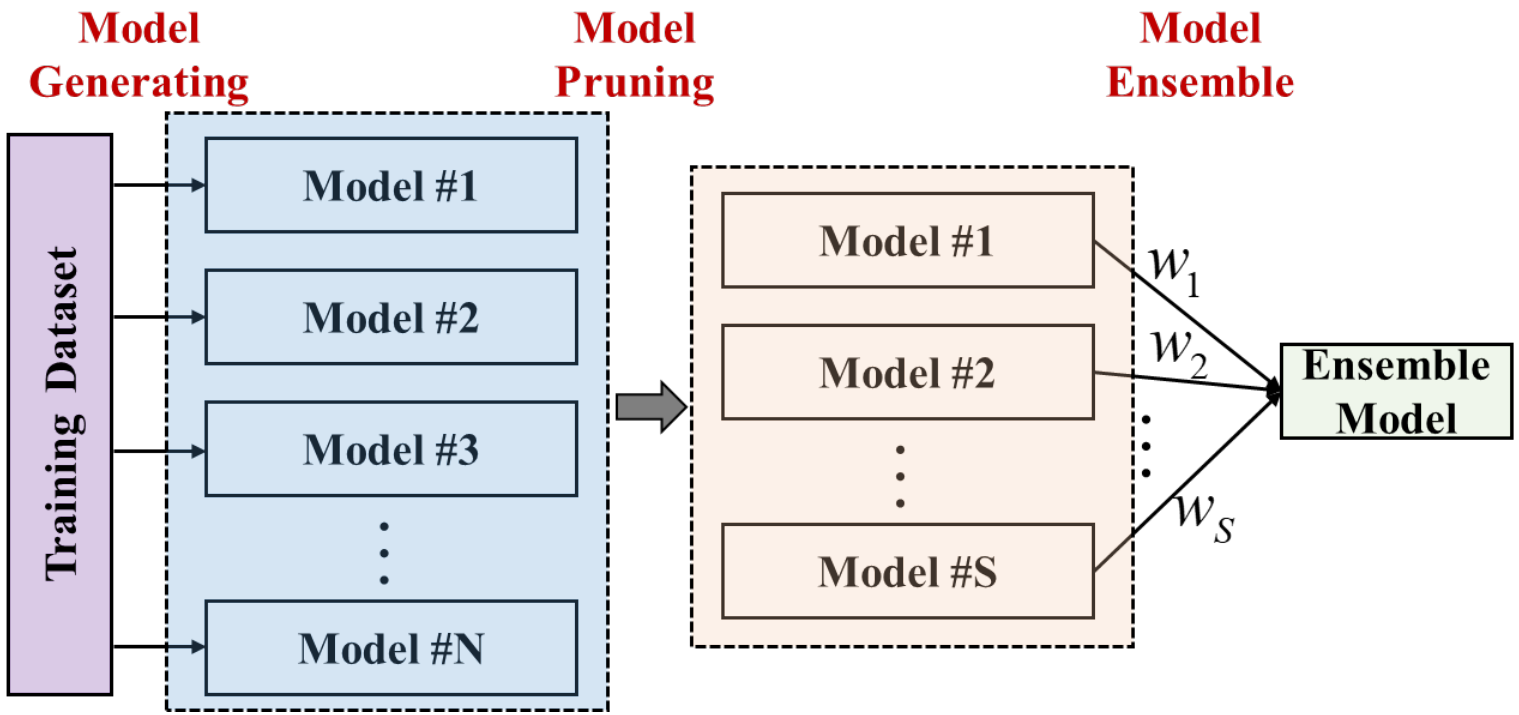
Combine point forecasts	Combine probabilistic forecasts
One dimension	High dimension
RMSE, MAPE	Reliability, sharpness, calibration
Analytical solution	???

Contributions of our work:

- ***New problem***: *Extend the ensemble method to the PLF area;*
- ***Elegant formulation***: *Formulate the combining problem into an LP or QP model.*

Combining probabilistic forecasts

Problem Formulation



1)How to generate multiple PLF models?

2)Among the *N* forecasting models, how many and which methods should be selected for the final ensemble formation process?

3) How much weight should be given to each method for the optimal combination?

Combining probabilistic forecasts

Problem Formulation

Loss function

Combined forecast

Determine the weights

$$\min_{\omega_n} \quad TL = \sum_{t=1}^{T} L\left(\sum_{n=1}^{N} \omega_n F_{n,t}, y_t\right)$$

Real load

$$\text{s.t.} \quad \sum_{n=1}^{N} \omega_n = 1, \quad \omega_n \geq 0$$

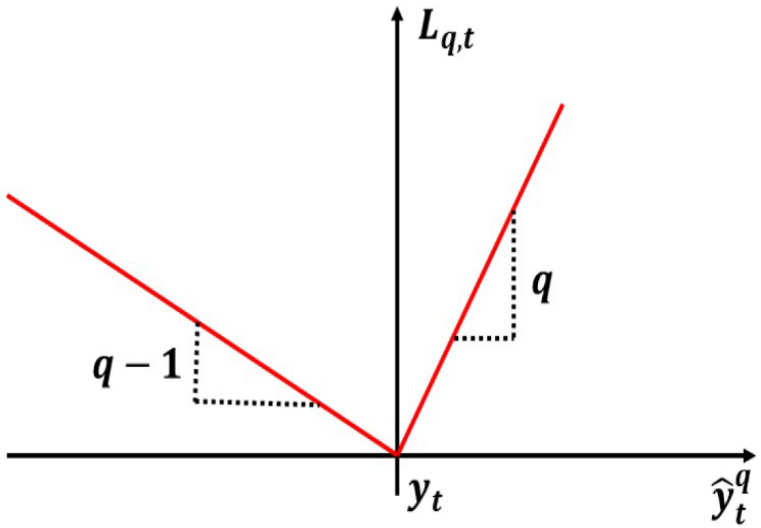
Summation and non-negative constraints

A deep investigation of the loss function is the key to formulate and solve the optimization problem.

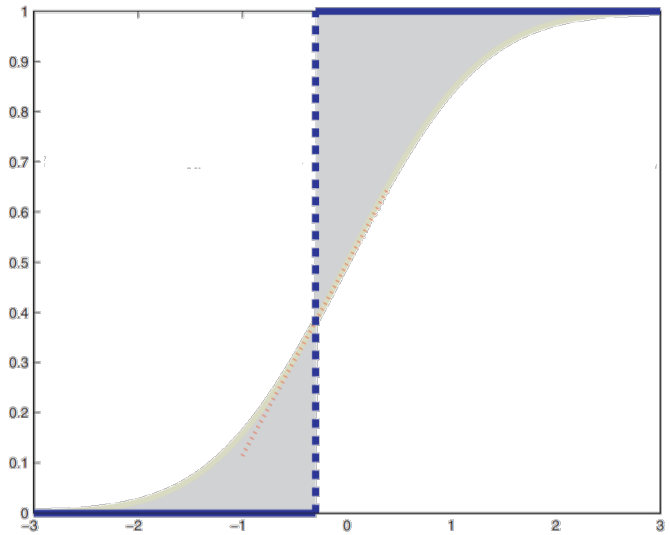
Combining probabilistic forecasts

Problem Formulation

Pinball loss *and* ***CRPS*** *assess the calibration and sharpness simultaneously, thus balancing the statistical consistency between the distributional forecasts and the observations and the concentration of the predictive distributions*



Pinball loss



Continuous ranked probability score (CRPS)

$$\mathrm{PL}\left(\hat{y}_{t,q}, y_t\right) = \begin{cases} \left(y_t - \hat{y}_{t,q}\right)q & \hat{y}_{t,q} \le y_t \\ \left(\hat{y}_{t,q} - y_t\right)(1-q) & \hat{y}_{t,q} > y_t \end{cases}$$

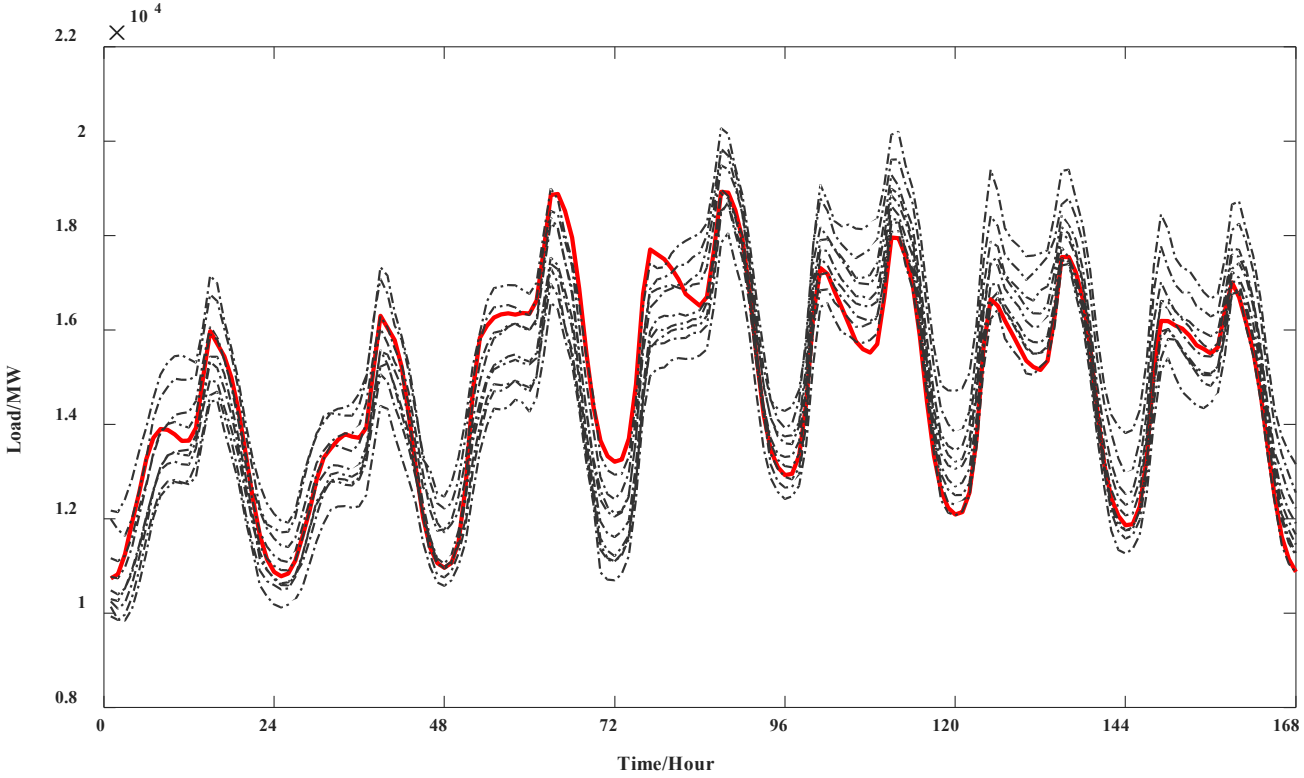
$$\mathrm{CRPS}\left(F_t, y_t\right) = \int_{-\infty}^{\infty} \left(F_t(z) - 1\left(z - y_t\right)\right)^2 dz$$

Combining probabilistic forecasts

Combining quantile forecasts

$$\mathrm{PL}\left(\hat{y}_{t,q}, y_t\right) = \begin{cases} \left(y_t - \hat{y}_{t,q}\right)q & \hat{y}_{t,q} \le y_t \\ \left(\hat{y}_{t,q} - y_t\right)(1-q) & \hat{y}_{t,q} > y_t \end{cases}$$

$$\hat{y}_{q,t} \approx \sum_{n=1}^{N} \omega_{n,q} \hat{y}_{n,q,t}$$



$$\min_{\omega_n} \quad TL = \sum_{t=1}^{T} L\left(\sum_{n=1}^{N} \omega_{n,q} \hat{y}_{n,q,t}, y_t\right)$$

$$s.t. \quad \sum_{n=1}^{N} \omega_{n,q} = 1, \omega_{n,q} \ge 0$$

Combining probabilistic forecasts

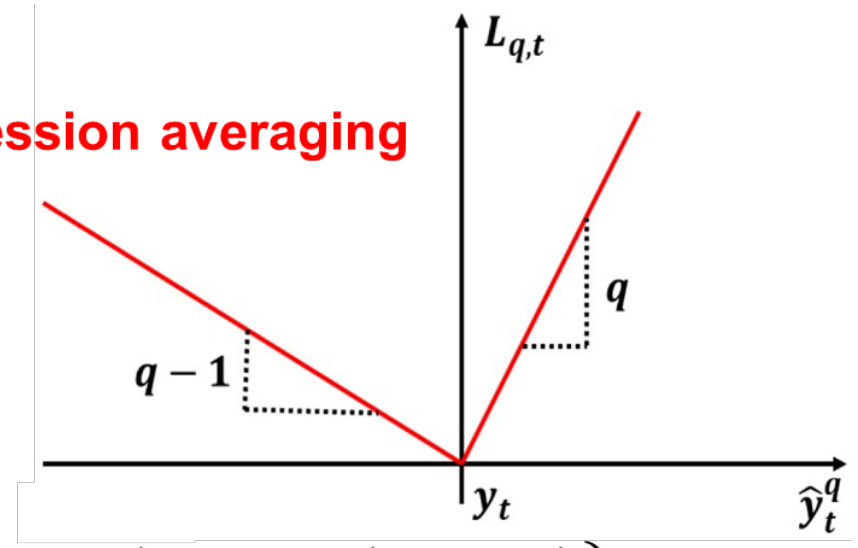
Combining quantile forecasts

$$\hat{\omega}_q = \arg\min_{\omega_q} \sum_{t\in T} L_{t,q}\left(\hat{y}_{t,q}, y_t\right)$$

Constrained quantile regression averaging

$$= \arg\min_{\omega_q} \sum_{t\in T} \max\left\{q\left(y_t - \hat{y}_{t,q}\right), (1-q)\left(\hat{y}_{t,q} - y_t\right)\right\}$$

$$\text{s.t. } \hat{y}_{q,t} \approx \sum_{n\in N} \omega_{n,q}\hat{y}_{n,t,q}, \quad \sum_{n\in N} \omega_{n,q} = 1, \quad \omega_{n,q} = 0 \quad \forall n.$$



Auxiliary decision variables $v_{t,q} = \max\left\{q\left(y_t - \hat{y}_{t,q}\right), (1-q)\left(\hat{y}_{t,q} - y_t\right)\right\}$

$$\hat{\omega}_q = \arg\min_{\omega_q} \sum_{t\in T} v_{t,q}$$

$$\text{s.t. } \hat{y}_{q,t} \approx \sum_{n\in N} \omega_{n,q}\hat{y}_{n,t,q}, \quad \sum_{n\in N} \omega_{n,q} = 1, \quad \omega_{n,q} = 0 \quad \forall n.$$

$$v_{t,q} \geq q\left(y_t - \hat{y}_{t,q}\right), \quad v_{t,q} \geq (1-q)\left(\hat{y}_{t,q} - y_t\right)$$

- **LP problem**
- **Model selection**

Combining probabilistic forecasts

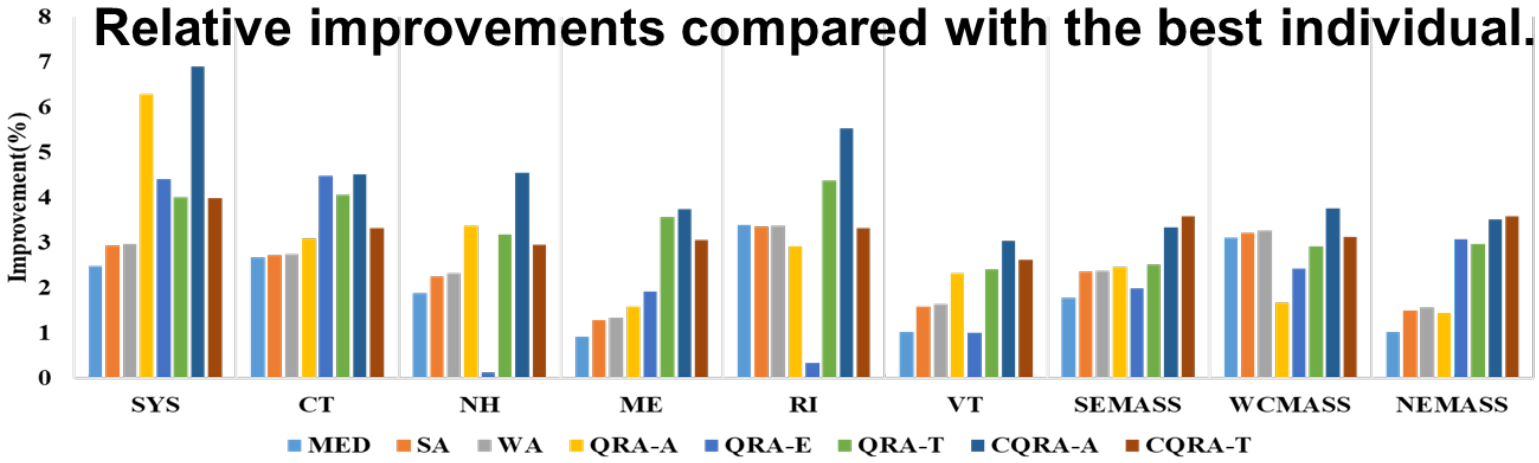
Combining quantile forecasts

We try to combine 13 individual methods and test the combined forecasts in ISO-NE dataset.

Pinball losses of different combining methods.

Methods \ Zones	SYS	CT	NH	ME	RI	VT	SEMASS	WCMASS	NEMASS
BI	288.563	81.478	27.216	18.146	21.756	12.426	42.307	41.939	63.685
NS	327.569	95.058	31.586	19.003	25.738	13.247	48.817	47.041	71.873
MED	281.607	79.359	26.713	17.981	21.044	12.300	41.570	40.676	63.048
SA	280.375	79.322	26.618	17.916	21.053	12.233	41.336	40.638	62.752
WA	280.266	79.306	26.600	17.908	21.049	12.227	41.329	40.616	62.706
QRA-E	276.417	77.995	27.184	17.806	21.683	12.303	41.484	40.949	61.793
QRA-A	271.519	79.037	26.330	17.864	21.140	12.145	41.295	41.252	62.783
QRA-T	277.487	78.313	26.380	17.523	20.847	12.135	41.271	40.752	61.849
CQRA-E	356.527	100.925	33.829	22.767	26.540	15.616	51.765	51.544	79.131
CQRA-A	277.510	78.870	26.437	17.610	21.059	12.109	**40.847**	40.672	**61.491**
CQRA-T	**269.953**	**77.961**	**26.034**	**17.492**	**20.619**	**12.061**	40.941	**40.422**	61.524

Relative improvements compared with the best individual.



Combining probabilistic forecasts

Combining quantile forecasts

Models that are selected for different quantiles for total load (SYS).

Models \ Quantiles	10-th	20-th	30-th	40-th	50-th	60-th	70-th	80-th	90-th
#1	0	0	0	0.128	0.123	0	0.015	0	0.102
#2	0	0	0	0.177	0.022	0.236	0.154	0.004	0
#3	0.036	0	0	0.041	0.255	0	0.123	0.302	0
#4	0.385	0.444	0.281	0	0	0.030	0	0	0.068
#5	0.165	0	0	0.200	0.298	0.339	0.092	0	0.134
#6	0.037	0.093	0.537	0.264	0	0	0.000	0.251	0
#7	0	0.131	0	0.071	0	0	0.265	0.051	0.218
#8	0	0.207	0.152	0	0.158	0.003	0.350	0.133	0
#9	0.377	0.047	0.030	0.117	0.143	0.392	0	0.206	0.333
#10	0	0.078	0	0	0	0	0	0	0
#11	0	0	0	0	0	0	0	0.052	0.145
#12	0	0	0	0	0	0	0	0	0
#13	0	0	0	0	0	0	0	0	0

Combining probabilistic forecasts

Combining quantile forecasts

Models that are selected for the 90-th quantile for different zones.

Models \ Zones	SYS	CT	NH	ME	RI	VT	SEMASS	WCMASS	NEMASS
#1	0.102	0.144	0.231	0.015	0.001	0.355	0	0	0.196
#2	0	0	0	0.082	0.074	0.146	0.071	0	0
#3	0	0	0.031	0	0	0.079	0	0.196	0
#4	0.068	0	0.089	0.349	0	0	0.038	0	0
#5	0.134	0	0	0	0.272	0	0.199	0.318	0.199
#6	0	0	0.283	0.231	0.226	0.096	0	0	0.136
#7	0.218	0	0.058	0.058	0	0.082	0.166	0.218	0.049
#8	0	0.129	0.308	0.079	0.197	0	0.173	0.076	0.087
#9	0.333	0.341	0	0.185	0.021	0.243	0.290	0.192	0.333
#10	0	0	0	0	0	0	0	0	0
#11	0.145	0.267	0	0	0	0	0	0	0
#12	0	0	0	0	0.210	0	0	0	0
#13	0	0.119	0	0	0	0	0.062	0	0

Combining probabilistic forecasts

Combining density forecasts

The applications of the CRPS have been hampered by a lack of readily computable solutions to the integral:

$$\mathrm{CRPS}(F_t, y_t) = \int_{-\infty}^{\infty} \left(F_t(z) - 1(z - y_t)\right)^2 dz$$

This drawback is overcome by [1]:

$$\mathrm{CRPS}(F_t, y_t) = E\left|Y - y\right| - \frac{1}{2} E\left|Y - Y'\right|$$

Let's consider a simple case: **Gaussian Mixture Distribution**

L. Baringhaus and C. Franz, "On a new multivariate two-sample test," Journal of Multivariate Analysis, vol. 88, no. 1, pp. 190–206, 2004.

Combining probabilistic forecasts

Combining density forecasts

Two lemmas for Gaussian model: $\mathrm{CRPS}(F_t, y_t) = E|Y - y| - \frac{1}{2}E|Y - Y'|$

Lemma 1: *The expectation of an absolute value of a finite mixture distribution is the* ***weighted sum of the corresponding expectations*** *of absolute values of the components of the finite mixture distribution.*
If $X_1, X_2, \cdots X_N$ *are the* N *components of the finite mixture distribution* X *with weights* $\omega_1, \omega_2, \cdots \omega_N$ *, then* $E|X| = \sum_{n=1}^{N} \omega_n E|X_n|$

Combining probabilistic forecasts

Combining density forecasts

Two lemmas for Gaussian model: $\mathrm{CRPS}(F_t, y_t) = E|Y - y| - \frac{1}{2}E|Y - Y'|$

Lemma 2: *If X and Y are independent random variables that are finite mixtures of normal distributions, then their sum is also a finite mixture of normal distributions. i.e., if*

$$f_X(x) = \sum_{l=1}^{L} \omega_l \phi(x \mid \mu_l, \sigma_l),\ f_Y(y) = \sum_{m=1}^{M} \omega_m \phi(y \mid \mu_m, \sigma_m)$$

$$\omega_l \geq 0,\quad \omega_m \geq 0,\quad \sum_{l=1}^{L} \omega_l = 1,\quad \sum_{m=1}^{M} \omega_m = 1$$

where $\phi(\cdot \mid \mu, \sigma)$ is the PDF of normal distribution $N(\mu, \sigma)$, then the PDF of $Z = X + Y$ is:

$$f_Z(z) = \sum_{l=1}^{L} \sum_{m=1}^{M} \omega_l \omega_m \phi(x \mid \mu_l + \mu_m, \sqrt{\sigma_l^2 + \sigma_m^2})$$

Combining probabilistic forecasts

Combining density forecasts

If individual density forecasts are **Gaussian distributed** $f_n(x)$, the combined forecast follows Gaussian mixture distribution:

$$f_X(x) = \sum_{n=1}^{N} \omega_n f_n(x)$$

Then, the CRPS can be calculated as:

$$\mathrm{CRPS}(F, y) = \sum_{n=1}^{N} \omega_n E\left|Y_n - y\right| - \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{1}{2}\omega_i \omega_j E\left|Y_i - Y_j'\right|$$

The expectation of the absolute value of a normal distribution $N(\mu,\sigma)$ can be calculated as follows:

$$E|X| = \int_{-\infty}^{\infty} |x| f(x)dx = \int_{-\infty}^{0} -xf(x)dx + \int_{0}^{\infty} |x| f(x)dx$$

$$= \sigma\sqrt{\frac{2}{\pi}} e^{-\frac{\mu^2}{2\sigma^2}} + \mu[2\Phi(\frac{\mu}{\sigma}) - 1]$$

Combining probabilistic forecasts

Combining density forecasts

Thus, we have:

$$\mathrm{CRPS}(F, y) = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_{i,j}\omega_i\omega_j + \sum_{n=1}^{N} \beta_n \omega_n$$

where

$$\alpha_{i,j} = \frac{1}{\sqrt{2\pi}}\sqrt{\sigma_i^2+\sigma_j^2}\exp\left(-\frac{(\mu_i-\mu_j)^2}{2(\sigma_i^2+\sigma_j^2)}\right) - \frac{\mu_i-\mu_j}{2}\left[2\Phi\left(\frac{(\mu_i-\mu_j)}{\sqrt{\sigma_i^2+\sigma_j^2}}\right)-1\right]$$

$$\beta_n = \sqrt{\frac{2}{\pi}}\sigma_n \exp\left(-\frac{(\mu_n-y)^2}{2\sigma_n^2}\right) + (\mu_n - y)\left[2\Phi\left(\frac{(\mu_n-y)}{\sigma_n}\right)-1\right]$$

Finally, we have:

$$\min_{\omega} \quad \omega^T Q \omega + c^T \omega$$

$$\text{s.t.} \quad \mathbf{1}^T \omega = 1 \quad \omega \geq 0$$

QP problem!
Is this convex?

Combining probabilistic forecasts

Combining density forecasts

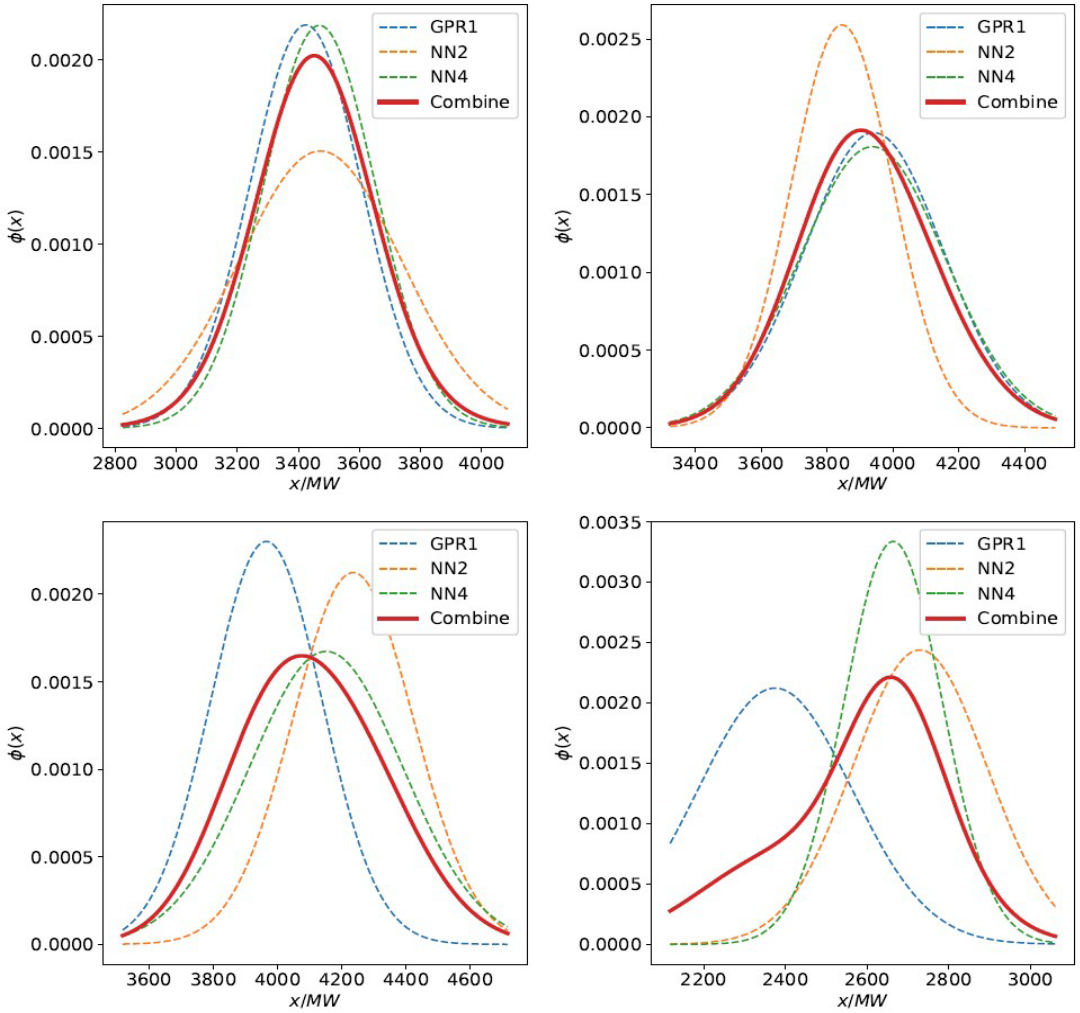
Performances of combined models

Metric \ Method	Simple Average	MAPE Based	CRPS Based
MAPE / %	5.48	5.09	5.11
MAE / MW	214.41	198.99	199.43
CRPS / MW	151.19	141.58	141.07

CRPS of the Best Individual Model and Combined Models

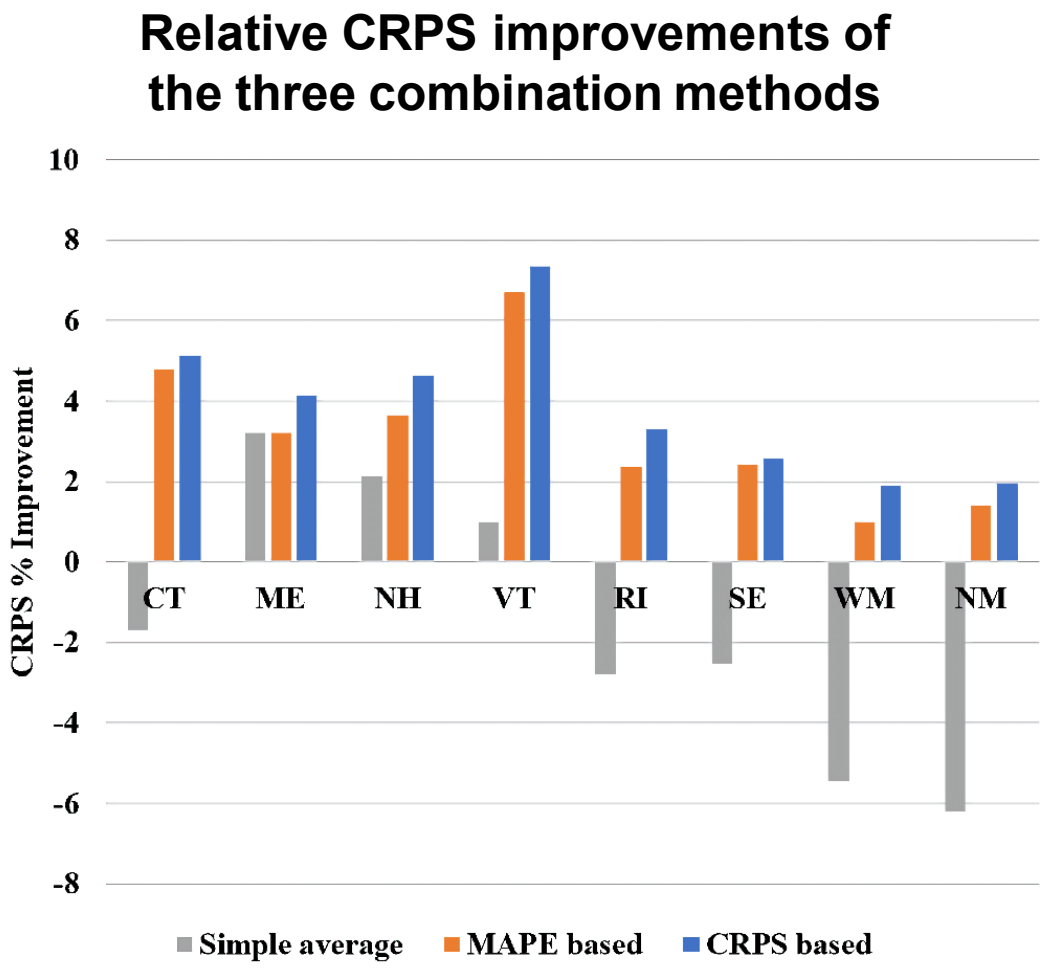
Region \ Method	Best Individual	Simple Average	MAPE Based	CRPS Based
CT	148.7	151.19	141.58	141.07
ME	33.8	32.71	32.71	32.40
NH	43.0	42.08	41.43	41.01
VT	21.2	20.99	19.78	19.64
RI	31.4	32.28	30.66	30.36
SE	61.4	62.95	59.91	59.83
WM	71.4	75.28	70.68	70.04
NM	103.7	110.14	102.25	101.66

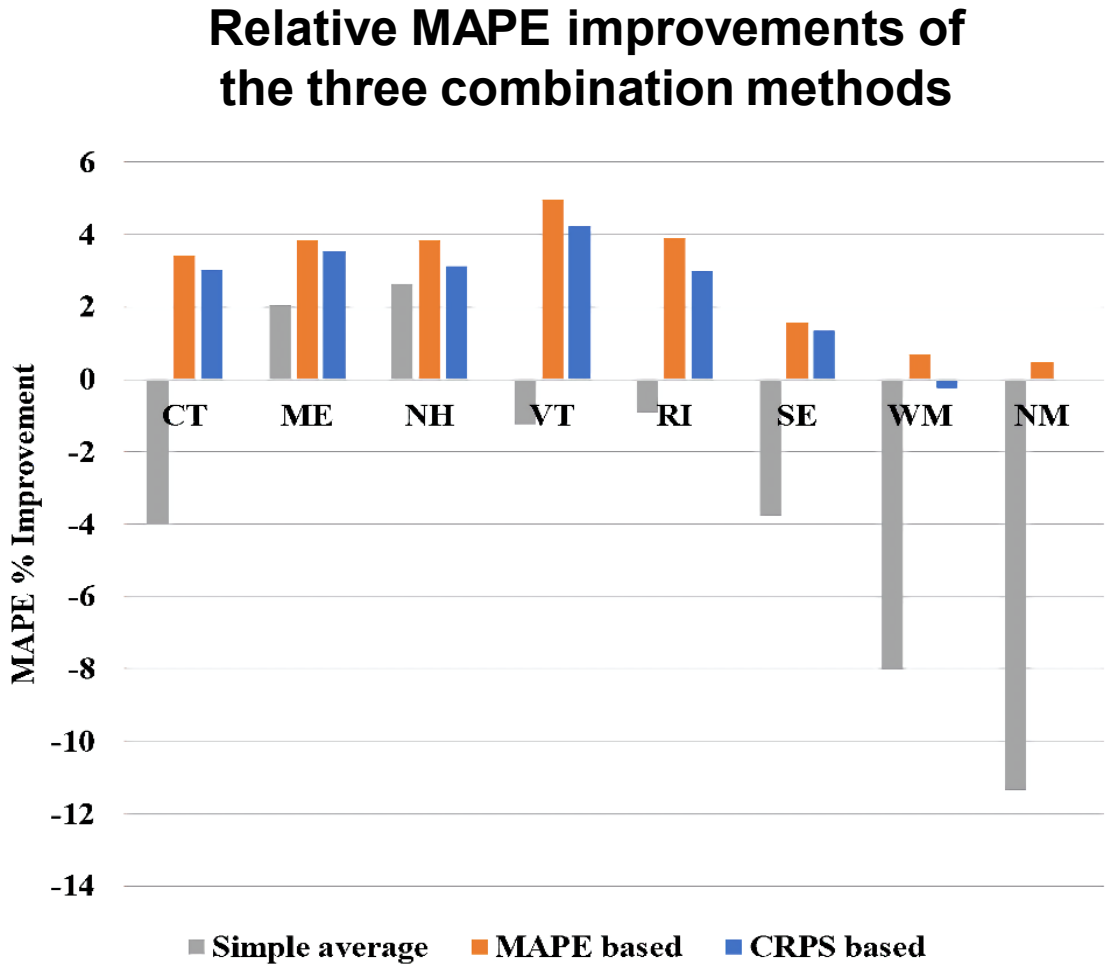
PDFs of predictions of four typical days given by the base models and their combination



Combining probabilistic forecasts

Combining density forecasts





Combining probabilistic forecasts

Combining density forecasts

Weights of the base models in the CRPS-guided model

	GPR1	GPR2	GPR3	GPR4	GPR5	NN1	NN2	NN3	NN4	NN5	XGB1	XGB2	XGB3	XGB4	XGB5
CT	0.33	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.47	0.00	0.00	0.00	0.00	0.00	0.00
ME	0.00	0.04	0.00	0.11	0.22	0.14	0.09	0.03	0.00	0.21	0.16	0.00	0.00	0.00	0.00
NH	0.15	0.01	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.07	0.09	0.18	0.14	0.06	0.27
VT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.12	0.07	0.12	0.14	0.13	0.23
RI	0.00	0.00	0.01	0.00	0.00	0.00	0.09	0.00	0.00	0.02	0.34	0.00	0.13	0.24	0.17
SEMASS	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.37	0.01	0.34
WCMASS	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.09	0.09	0.00	0.38	0.13	0.22
NEMASSBOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.04	0.00	0.31	0.33	0.15

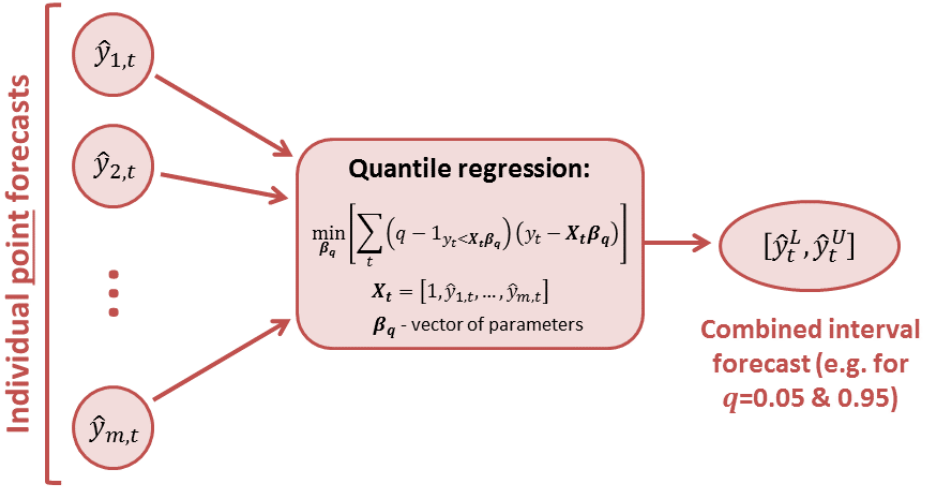
Weights of the base models in the MAPE-guided model

	GPR1	GPR2	GPR3	GPR4	GPR5	NN1	NN2	NN3	NN4	NN5	XGB1	XGB2	XGB3	XGB4	XGB5
CT	0.26	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.63	0.00	0.00	0.00	0.00	0.00	0.00
ME	0.00	0.12	0.00	0.00	0.40	0.00	0.32	0.02	0.00	0.06	0.07	0.00	0.00	0.00	0.00
NH	0.08	0.16	0.01	0.00	0.00	0.11	0.00	0.03	0.00	0.02	0.22	0.00	0.00	0.10	0.27
VT	0.00	0.00	0.04	0.00	0.00	0.02	0.00	0.00	0.15	0.05	0.05	0.13	0.29	0.00	0.26
RI	0.00	0.02	0.03	0.00	0.08	0.00	0.10	0.04	0.00	0.00	0.30	0.00	0.31	0.13	0.00
SEMASS	0.00	0.08	0.01	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.10	0.00	0.48	0.00	0.29
WCMASS	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.12	0.20	0.00	0.60	0.00	0.00
NEMASSBOST	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.54	0.43	0.00

Combining probabilistic forecasts

Quantile Regression Averaging

Quantile Regression Averaging (QRA) is a forecast combination approach to the computation of prediction intervals. It involves applying **quantile regression** to the point forecasts of a small number of individual forecasting models or experts.



1. **Generate individual point forecasts;**
2. **Quantile regression;**
3. **Final forecasts.**

Nowotarski, Jakub; Weron, Rafał (2015). "Computing electricity spot price prediction intervals using quantile regression and forecast averaging". *Computational Statistics*. 30 (3): 791–803.

Combining probabilistic forecasts

Generate individual point forecasts

Sister recency effect load forecasting models:

$$\hat{y}_t = \beta_0 + \beta_1 M_t + \beta_2 W_t + \beta_3 H_t + \beta_4 W_t H_t + f(T_t) + \sum_d f(\tilde{T}_{t,d}) + \sum_{\text{lag}} f(\tilde{T}_{t-\text{lag}})$$

where:

$$f(T_t) = \beta_5 T_t + \beta_6 T_t^2 + \beta_7 T_t^3 + \beta_8 T_t M_t + \beta_9 T_t^2 M_t + \beta_{10} T_t^3 M_t + \beta_{11} T_t H_t + \beta_{12} T_t^2 H_t + \beta_{13} T_t^3 H_t$$

$$\tilde{T}_{t,d} = \frac{1}{24} \sum_{\text{lag}=24d-23}^{24d} T_{t-\text{lag}}$$

By tuning the length of the training dataset and the partition of the training and validation datasets for model selection, we can obtain different “average-lag” (or d-lag) pairs, leading to different sister models.

B. Liu, J. Nowotarski, T. Hong and R. Weron, "Probabilistic Load Forecasting via Quantile Regression Averaging on Sister Forecasts," IEEE Transactions on Smart Grid, vol. 8, no. 2, pp. 730-737, March 2017.

Combining probabilistic forecasts

Performance on GEFCom2014

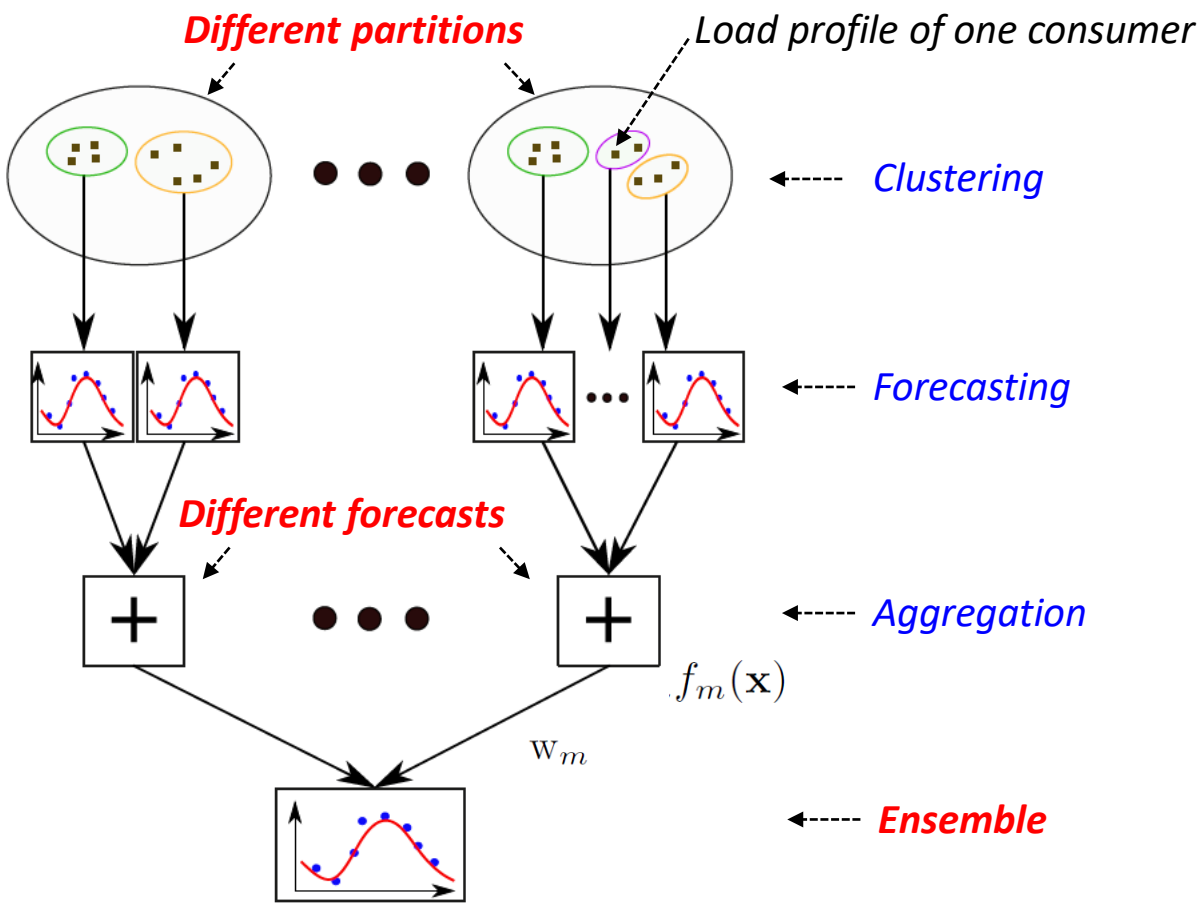
TABLE II
FORECASTING RESULTS COMPARISON IN THE TEST PERIOD (I.E., YEAR 2010) FOR THE MODELS SELECTED IN THE VALIDATION PERIOD (SEE TABLE I)

Model class	Pinball		Winkler (50%)		Winkler (90%)	
	(S,L)	Score	(S,L)	Score	(S,L)	Score
QRA	(8,183)	2.85	(7,183)	25.04	(7,365)	55.85
Ind	(1,91)	3.22	(1,91)	26.35	(1,365)	56.38
BI	(-,365)	3.00	(-,365)	26.38	(-,365)	57.17
Direct	-	3.19	-	26.62	-	94.27
Vanilla	-	8.00	-	70.51	-	150.0

B. Liu, J. Nowotarski, T. Hong and R. Weron, "Probabilistic Load Forecasting via Quantile Regression Averaging on Sister Forecasts," IEEE Transactions on Smart Grid, vol. 8, no. 2, pp. 730-737, March 2017.

Combining probabilistic forecasts

Generate individual point forecasts



If there are different partitions of consumers, we can obtain different bus load forecasts.

Combined model:

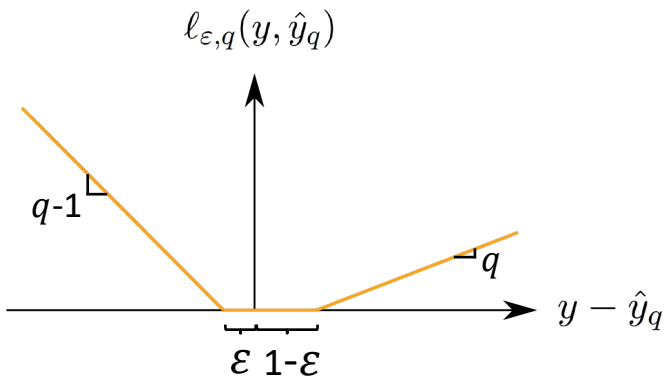
$$G(\mathbf{x}) = \sum_{m=1}^{M} \mathrm{w}_m f_m(\mathbf{x})$$

Combining probabilistic forecasts

Quantile Passive Aggressive Regression

General Formula

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left[\, d(\mathbf{w}, \mathbf{w}_t) + \eta_t \ell(y_t, \mathbf{w} \cdot \mathbf{x}_t) \, \right]$$



L_2-distance :

$$d(\cdot) = \frac{1}{2} \|\cdot\|^2$$

ε-insensitive quantile loss :

$$\ell_{\varepsilon,q}(\mathbf{w}_q; \mathbf{x}, y) = \begin{cases} q(y - \mathbf{w}_q \cdot \mathbf{x} + \varepsilon(q-1)) & \text{if } y - \mathbf{w}_q \cdot \mathbf{x} > \varepsilon(1-q) \\ 0 & \text{if } -\varepsilon q < y - \mathbf{w}_q \cdot \mathbf{x} < \varepsilon(1-q) \\ (q-1)(y - \mathbf{w}_q \cdot \mathbf{x} + \varepsilon q) & \text{if } y - \mathbf{w}_q \cdot \mathbf{x} < -\varepsilon q \end{cases}$$

Solving KKT conditions:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta_t sign(y_t - \mathbf{w}_t \cdot \mathbf{x}_t) \tau_t \mathbf{x}_t \qquad \tau_t = min \left\{ C, \frac{\ell_{\varepsilon,q}(y_t, \mathbf{w}_t \cdot \mathbf{x}_t)}{q \, \|\mathbf{x}_t\|_2^2} \right\}$$

Combining probabilistic forecasts

Performance on Irish Load Data

Errors on test set after batch learning

Method	ACE	PBL	WKS
QSGD	-0.92%	51.60	722.43
QPAR	2.23%	47.61	1075.02
QNN	-2.55%	54.94	776.86
Batch QRA	-5.25%	44.55	734.64
Window QRA	-1.90%	40.30	659.94

*QSGD: Quantile Stochastic Gradient Descent
*QPAR: Quantile Passive Aggressive Regression
*QNN: Quantile Neural Network

Errors on test set after online learning

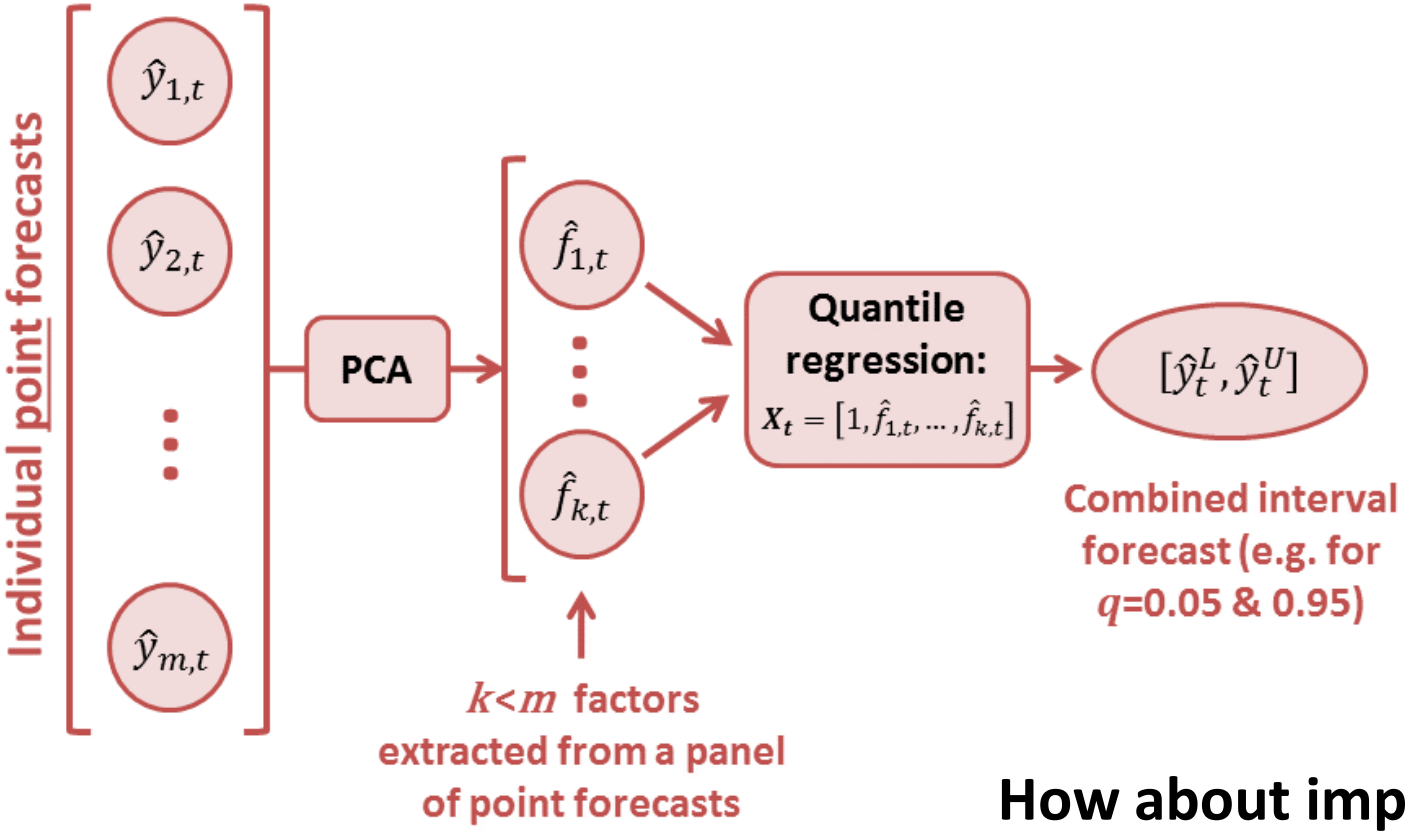
Method	ACE	PBL	WKS
QSGD	-0.02%	30.04	527.94
QPAR	-1.69%	29.47	484.59
QNN	-0.64%	56.10	930.23
Batch QRA	-5.25%	44.55	734.64
Window QRA	-1.90%	40.30	659.94

*Window OPT: window-based optimization

- All ensembles outperform the benchmarks after online learning except QNN
- The proposed method has the highest accuracy regarding pinball loss and winkler score
- A substantial performance improvement can be achieved by ensembles incorporating online learning.

Combining probabilistic forecasts

Quantile Regression Averaging



How about improvement on quantile regression?

Combining probabilistic forecasts

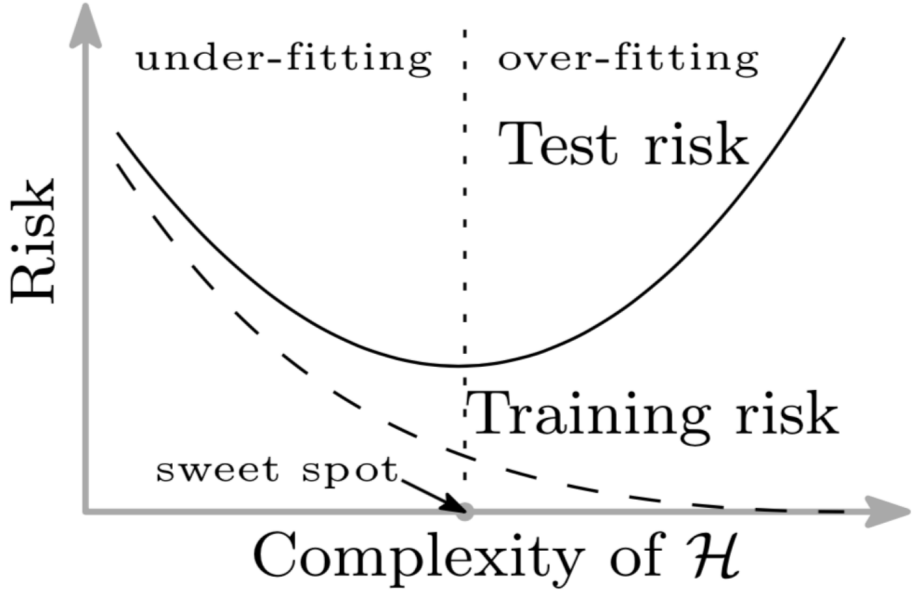
Summary:

- Procedure for combining probabilistic forecasts;
- Selection of individual forecasts with diversity;
- Combining quantile/density/sister forecasts;
- Data split is very important: train-validation-combine-test sets.

Data Sharing for probabilistic forecasting

Data Sharing for probabilistic forecasting

Why we need to share data?



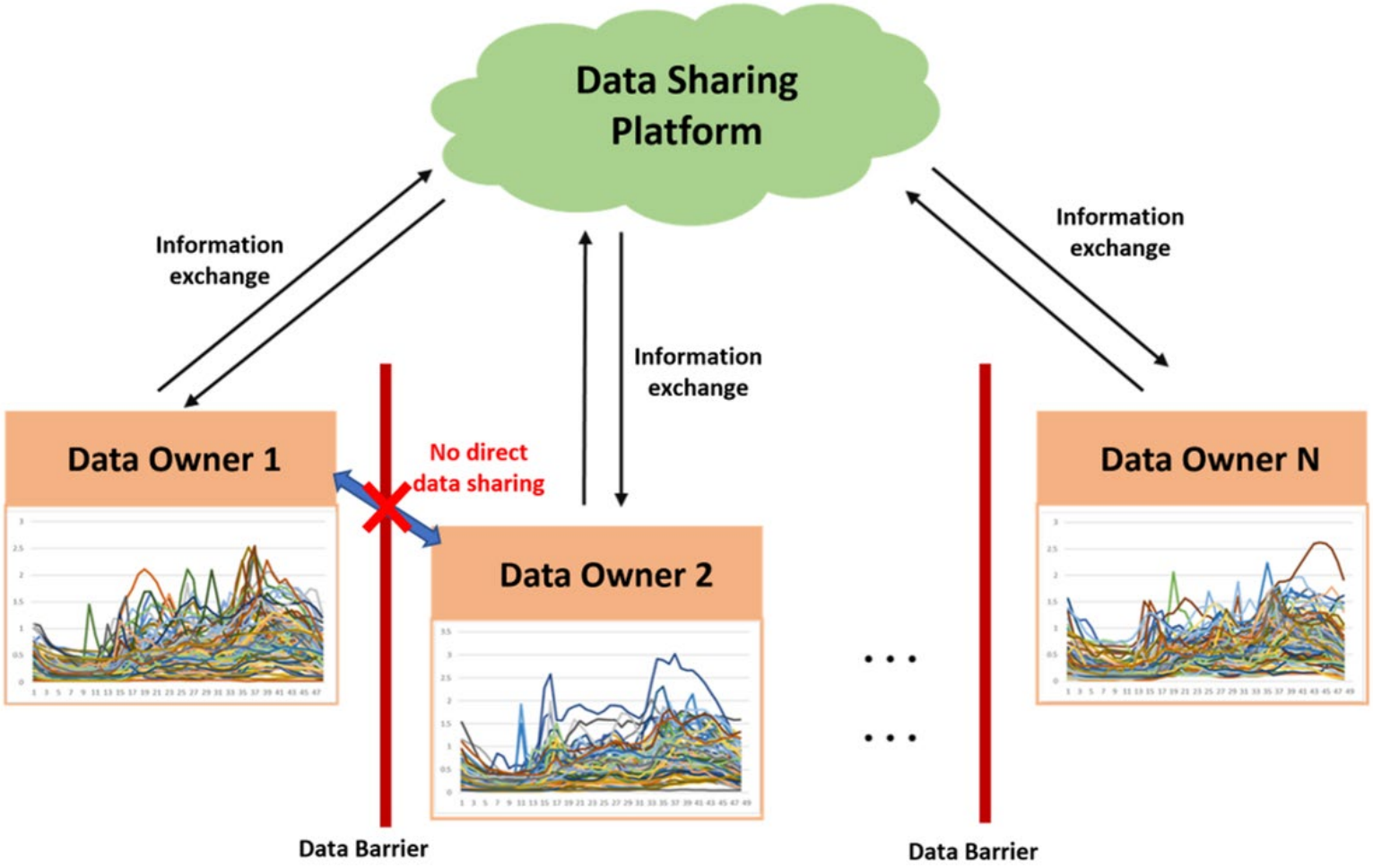
Reducing overfitting risk

Integrating more information

Mohri M, Muñoz Medina A. New analysis and algorithm for learning with drifting distributions[C]//International Conference on Algorithmic Learning Theory. Springer, Berlin, Heidelberg, 2012: 124-138.

Data Sharing for probabilistic forecasting

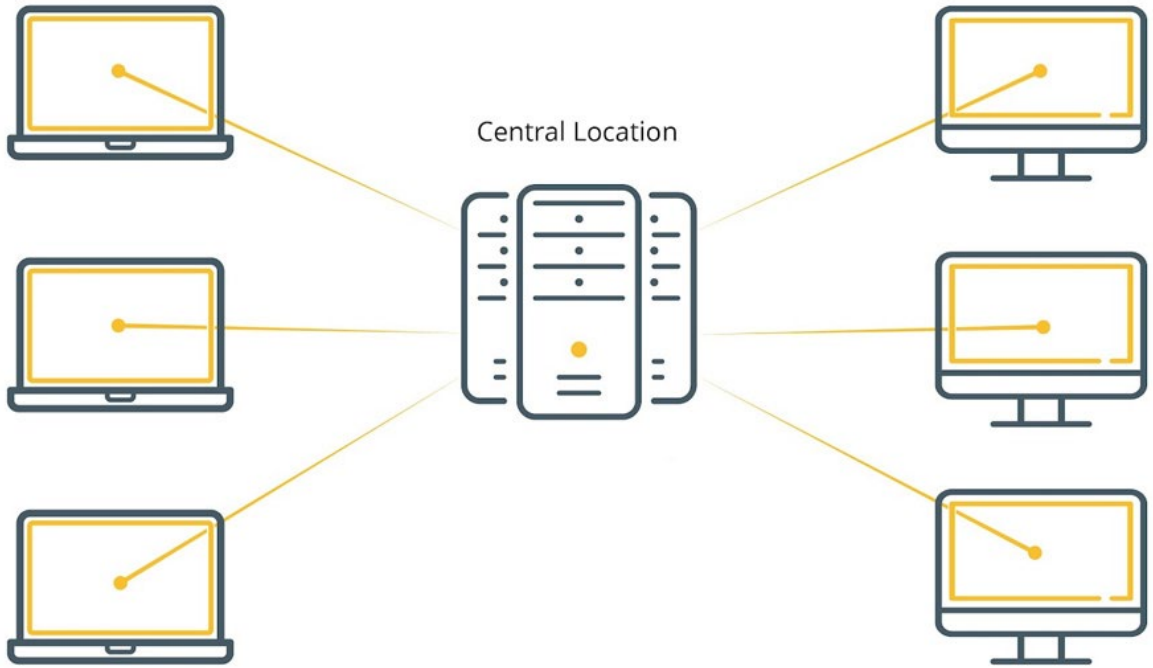
Privacy Issues in Forecasting



Data Sharing for probabilistic forecasting

Privacy Issues in Forecasting

Traditional ML-based forecasting algorithms assume that private data can be freely accessed from a centralized location.



Such centralized scheme brings privacy concerns and becomes impracticable for mainly two concrete reason:

- **Legislation for privacy protecting**
- **Commercial confidentiality of data in competitive environment.**

Data Sharing for probabilistic forecasting

Privacy Issues in Forecasting

Legislation for Privacy Protecting

- UK launched two policies in 2018, **Smart Meter Bill** and **The Data Protection Act**, respectively authorizing half-hourly electricity consumption data collection and implementing **General Data Protection Regulation (GDPR)** to utilize consumers' data and protect data privacy

- Canada implemented the **Personal Information Protection** and **Electronic Documents Act (PIPEDA)** in 2019, stating the main purpose that when smart meter data is gathered, used, or revealed, the individual's consent is required, and personal information can only be used for the stated specific purposes.

- To address privacy concerns with smart grid technology, the Office of Electricity Delivery and Energy Reliability and the Federal Smart Grid Task Force has published a **Voluntary Code of Conduct (VCC)** for utilities and third parties in the United States

- **Data Security Law of the Peoples Republic of China** was passed in 2021, which strictly regularizes data collection, storage, use, processing, transport, provision, and disclosure.

- ……

Data Sharing for probabilistic forecasting

Privacy Issues in Forecasting

Commercial Confidentiality of Data

In a open market, a competitive advantage can be gained through data analysis and personalized customer service; therefore, utilities may be hesitant to share their data in this circumstance.

Data Sharing for probabilistic forecasting

Privacy Issues in Forecasting

Problems & Solutions

How can high-performance forecasting be carried out in scenarios where data privacy is crucial?

- ***Privacy-preserving techniques***
- ***Collaborative forecasting***
- ***Federated forecasting***
- ***……***

Data Sharing for probabilistic forecasting

Privacy-Preserving Techniques

Six privacy-preserving techniques for sharing data

Fully homomorphic encryption: Data is encrypted before it is shared. It can be analyzed, but not decoded into the original information.

Differential privacy: Noise is added to the dataset so that it is impossible to reverse-engineer the original inputs.

Functional encryption: Select users have a key that allows them to view some parts of encrypted text.

Federated analysis: Parties share insights from their analysis without sharing the data itself.

Zero-knowledge proofs: Users can prove their knowledge of a value without revealing the value itself.

Secure multiparty computation: Data analysis is spread across multiple parties such that no single party can see the complete set of inputs.

Source: Deloitte research and analysis.

There is an inherent trade-off between the performance of the prediction model and the privacy of the local data.

Data Sharing for probabilistic forecasting

Collaborative Forecasting

Definition & Goal

Because of the spatio-temporal relationships in time series data, collaboration among several renewable energy power plant owners helps increase forecast skill.

Instead of sharing raw data, the model fitting problem is tackled in a distributed manner with collaborative forecasting, which observes following principles:

- *Improve forecast skill as compared to a scenario without collaboration*
- *Protect data privacy*
- *Consider synchronous and asynchronous communication between agents*

Data Sharing for probabilistic forecasting

Collaborative Forecasting

Schemes

Two collaborative schemes are possible: centralized communication with a central node (central hub) and peer-to-peer communication (P2P).

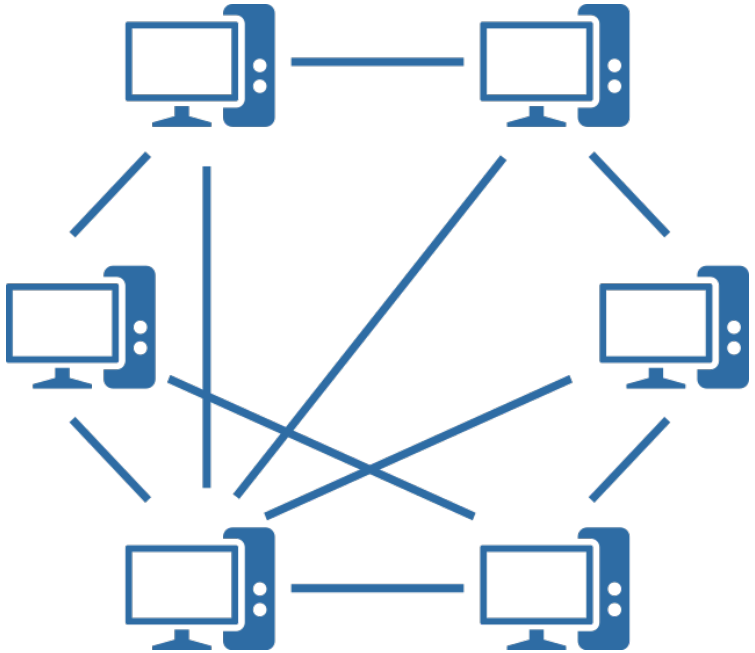
Central hub scheme:

- The scope of the calculations performed by the agents in the central hub model is restricted by their local data, only statistics data is transferred to the central node (such as average values or local data multiplied by locally determined coefficients).
- The central node combines these local estimators and coordinates the separate optimization processes to solve the main optimization problem.
- A transmission/distribution system operator (TSO/DSO) or a forecasting service provider can serve as the central node. The TSO or DSO can operate a platform that encourages collaboration amongst competitive RES power units to increase forecasting accuracy and lower system balancing costs.

Data Sharing for probabilistic forecasting

Collaborative Forecasting

P2P scheme:



- In the P2P model, the agents each compute their estimators locally but share their results with their peers, implying that each agent is both an agent and a central node.
- While P2P is more resilient (i.e., there are fewer points of failure), it is usually difficult to make it as cost-effective as the central hub model.
- The peer-to-peer paradigm is appropriate for data owners who do not wish to rely (or trust) on a neutral third party.
- P2P forecasting between prosumers or RES power plants could be potential business models, as could smart cities, which are characterized by a growing number of sensors and gadgets put at homes, buildings, and transit networks.

Data Sharing for probabilistic forecasting

Collaborative Forecasting

Example 1

Privacy-preserving LASSO-VAR

MEAN RELATIVE NRMSE IMPROVEMENT [%] OVER THE LASSO-AR MODEL

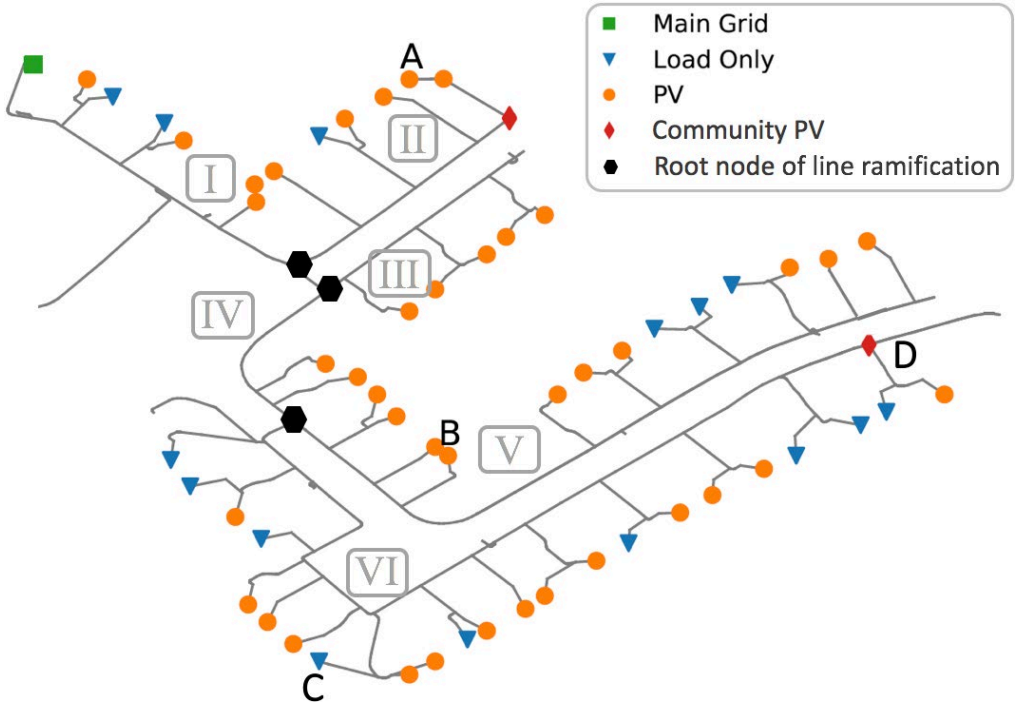
	h=1		h=2		h=3		h=4		h=5		h=6	
p_i	central	P2P	central	P2P	central	P2P	central	P2P	central	P2P	central	P2P
0	8.41		6.05		2.95		1.52		1.39		0.93	
0.1	7.93	8.41	5.98	6.05	2.91	2.95	1.49	1.52	1.35	1.39	0.89	0.93
0.3	7.45	”	5.89	”	2.89	”	1.40	”	1.18	”	0.69	”
0.5	6.69	”	5.77	”	2.88	”	1.30	”	1.00	”	0.52	”
0.7	5.71	”	5.54	”	2.84	”	1.24	”	0.89	”	0.33	”
0.9	3.75	8.10	5.19	5.75	2.74	2.78	0.75	1.47	0.62	1.38	-0.82	0.88

Goncalves C, Bessa R J, Pinson P. Privacy-preserving distributed learning for renewable energy forecasting[J]. IEEE Transactions on Sustainable Energy, 2021, 12(3): 1777-1787.

Data Sharing for probabilistic forecasting

Federated Forecasting

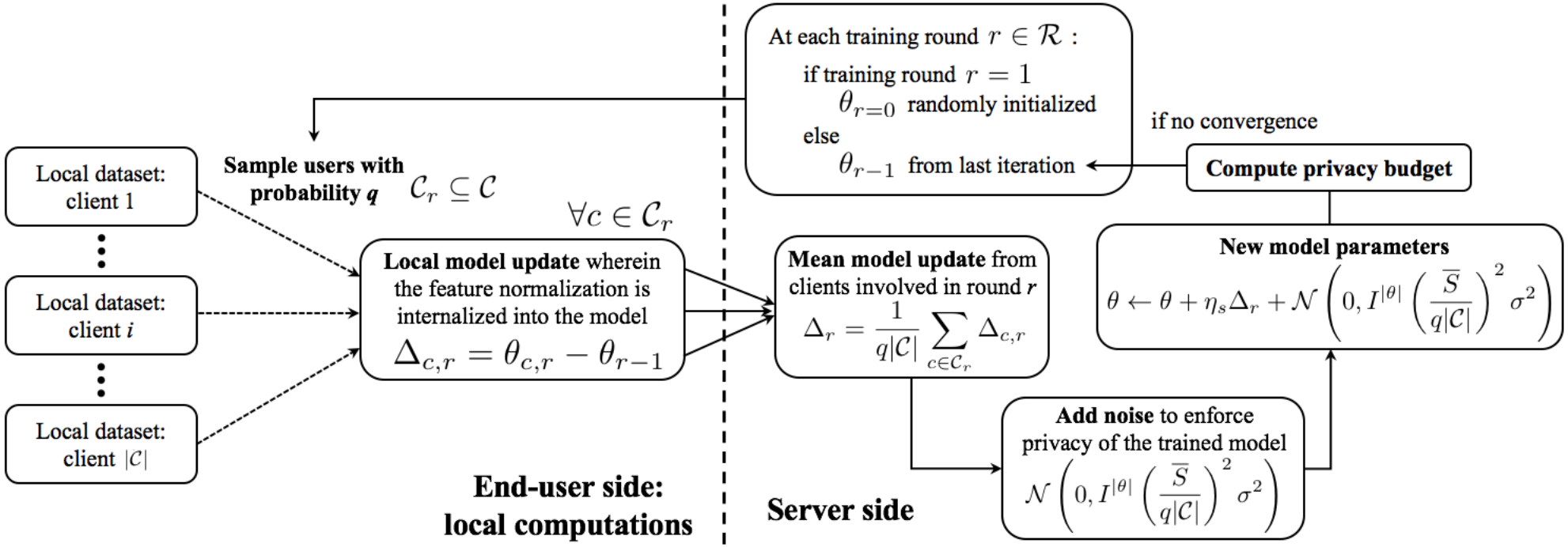
Privacy-Preserving Probabilistic Voltage Forecasting



Data Sharing for probabilistic forecasting

Federated Forecasting

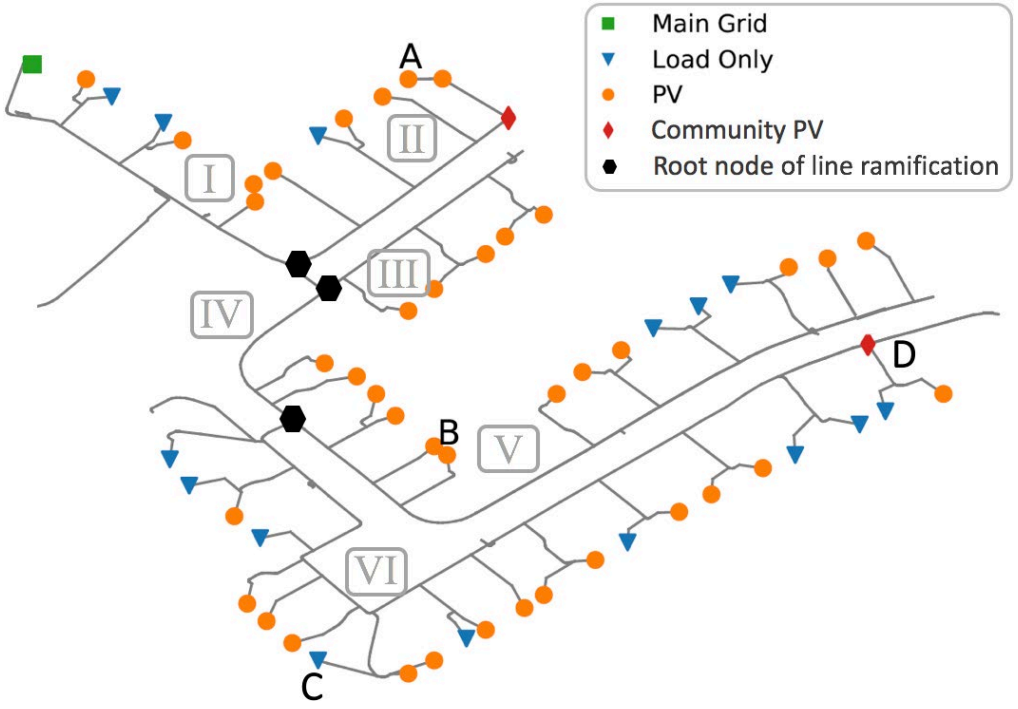
Privacy-Preserving Probabilistic Voltage Forecasting

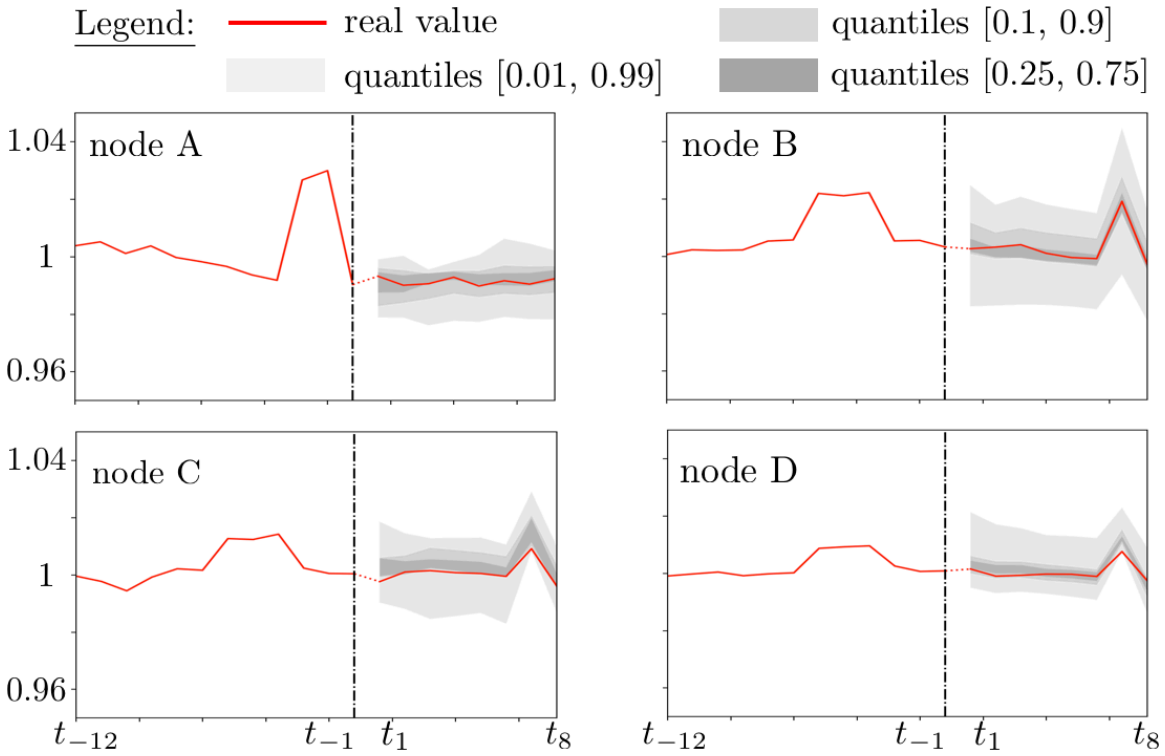


Data Sharing for probabilistic forecasting

Federated Forecasting

Privacy-Preserving Probabilistic Voltage Forecasting





Data Sharing for probabilistic forecasting

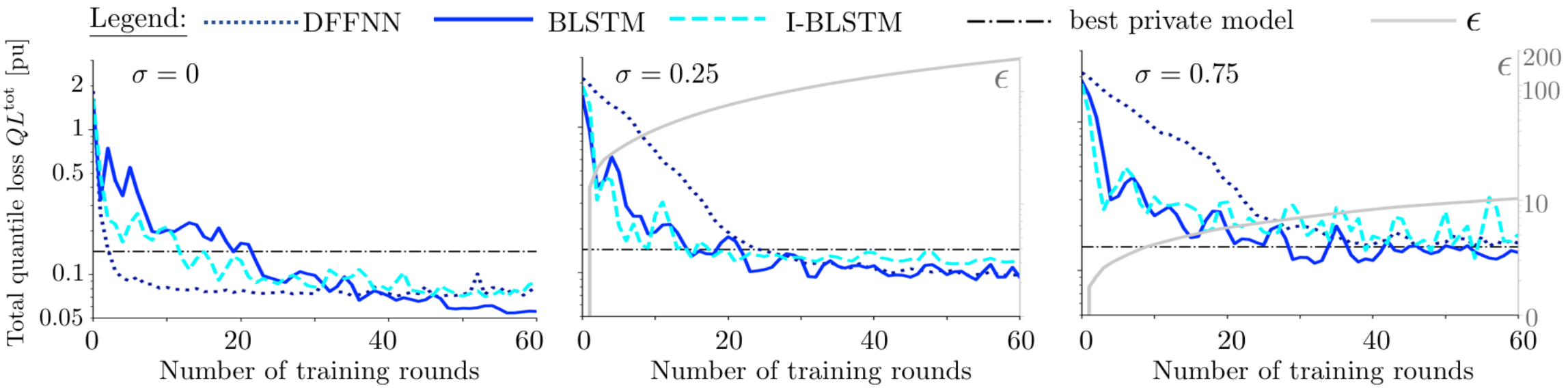
Federated Forecasting

PERFORMANCE AND TRAINING TIME OF DIFFERENT MODELS.

Model	Centralized		Local (private)	
	QL^{tot} [pu]	time [min] (epochs)	QL^{tot} [pu]	time [min] (epochs)
Prob-Persist	-	-	0.210	-
Prob-Avg	0.319	-	0.192	-
QRF	0.048	148	0.156	172
QGBDT	0.041	64	0.151	76
DFFNN	0.043	10 (33)	0.144	16 (27)
LSTM	0.034	52 (18)	0.186	63 (4)
BLSTM	0.032	88 (16)	0.172	105 (4)
A-BLSTM	0.192	54 (11)	-	-

Data Sharing for probabilistic forecasting

Federated Forecasting



Trade-off between privacy and performance

Data Sharing for probabilistic forecasting

Federated Forecasting

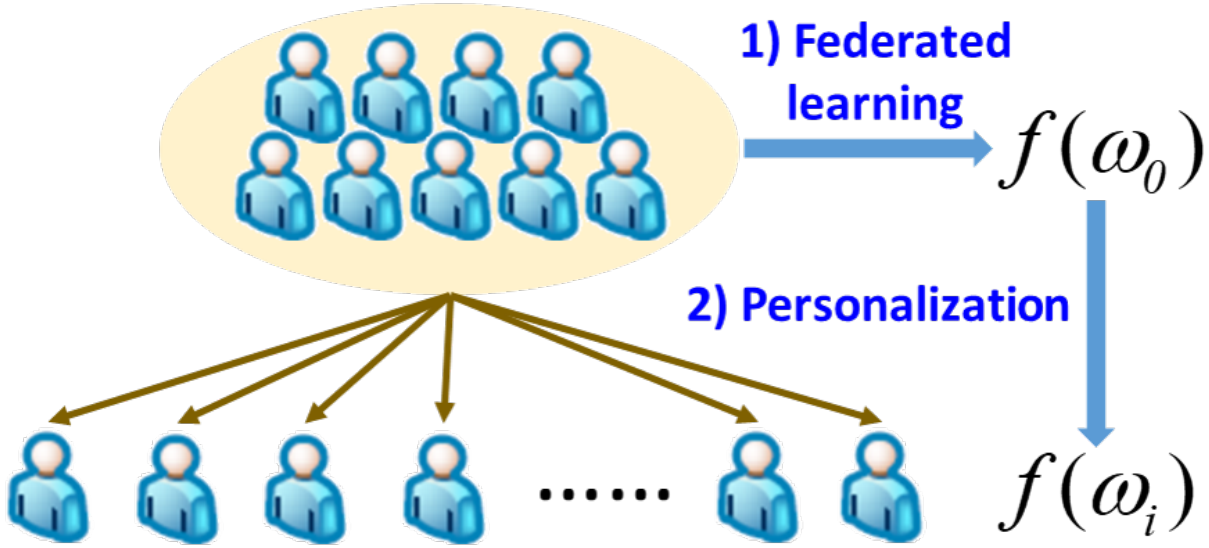
PERFORMANCE OF THE BLSTM FOR DIFFERENT PRIVACY SETTINGS.

Model	$q\|\mathcal{C}\|$	σ	r^{opt}	ϵ^{opt}	QL^{tot} [pu]
BLSTM	5	0	59	-	0.060
BLSTM	5	0.25	60	105.1	0.104
BLSTM	5	0.75	38	4.0	0.132
BLSTM	10	0	57	-	0.054
BLSTM	10	0.25	54	167.7	0.090
BLSTM	10	0.75	32	7.3	0.112

Data Sharing for probabilistic forecasting

Federated Forecasting

Personalized federated individual load forecasting



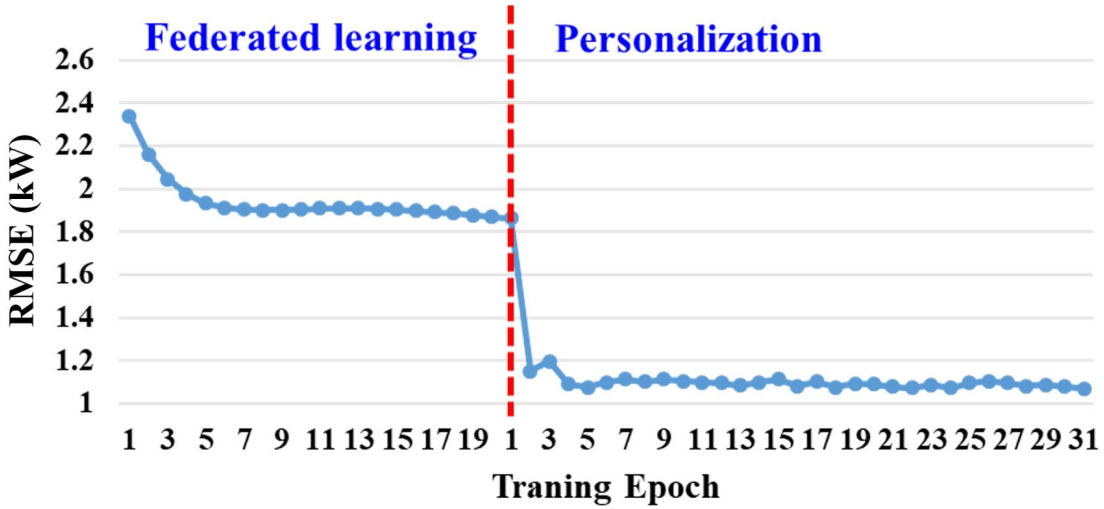
$$\min_{\omega_0} L(\omega_0) = \frac{1}{P}\sum_{n=1}^{N}\sum_{k=1}^{p_n}(f(\omega_0; x_{n,k}) - y_{n,k})^2$$

$$\arg\min_{\omega_n} \frac{1}{p_n}\sum_{k=1}^{p_n}(f(\omega_0; x_{n,k}) - y_{n,k})^2 + \frac{d_n\alpha}{\sqrt{p_n}}||\omega_n - \omega_0||^2$$

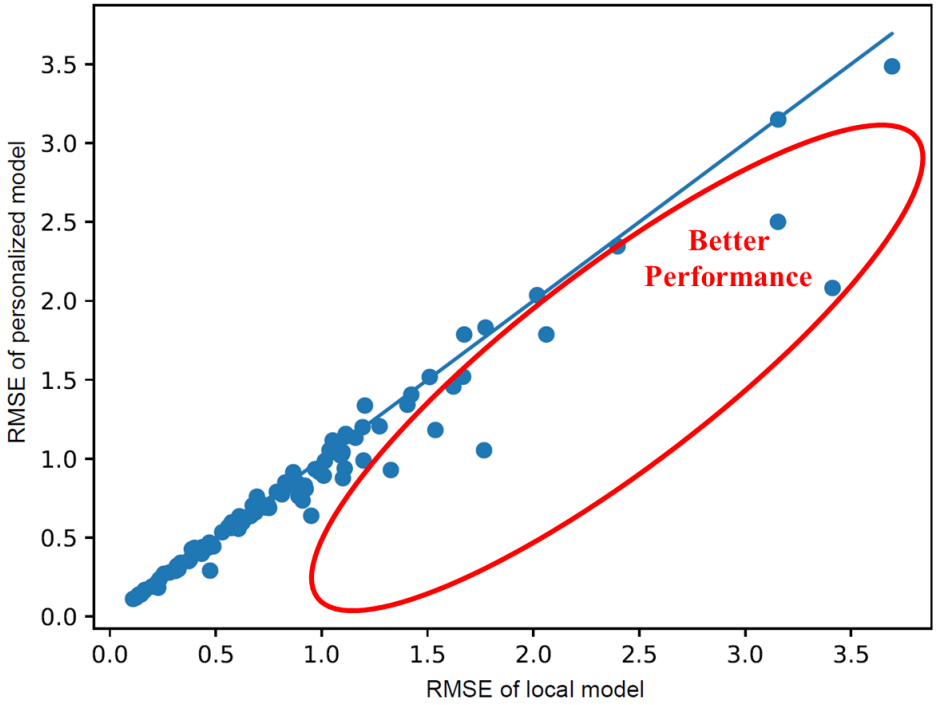
Data Sharing for probabilistic forecasting

Federated Forecasting

Personalized federated individual load forecasting



Changes of RMSE over client 1443 on training set.



Performance (RMSE) of personalized model against local model.

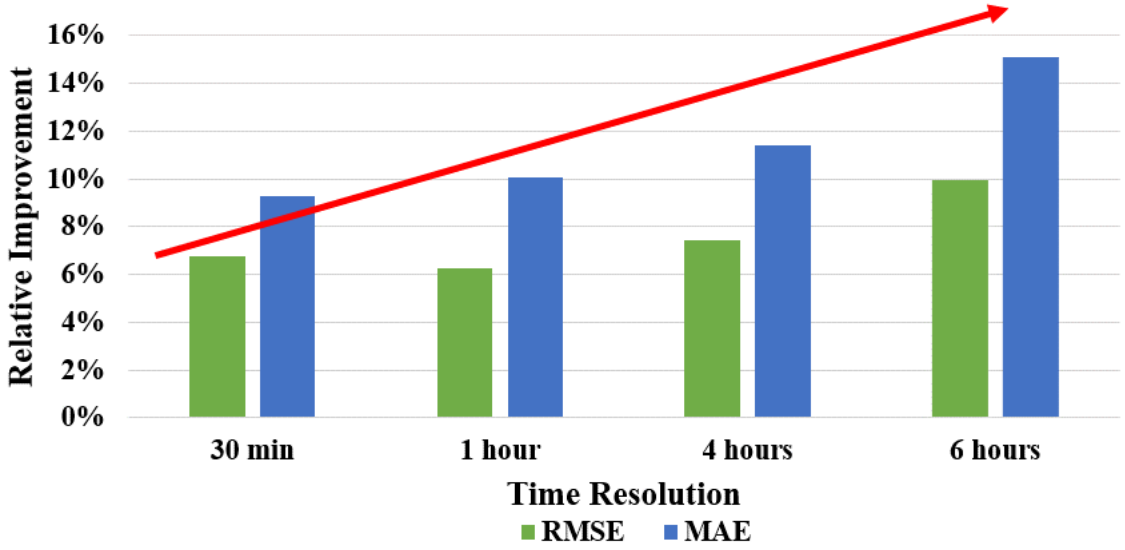
Data Sharing for probabilistic forecasting

Federated Forecasting

Personalized federated individual load forecasting

TABLE I: Performance with different resolutions (kW)

	Local		Federated		Personalized	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
30min	0.891	0.563	2.327	1.466	0.826	0.506
1 hour	1.776	1.134	3.616	2.351	1.650	1.009
4 hours	6.705	4.557	10.834	7.849	6.214	4.051
6 hours	9.087	6.400	14.835	11.080	8.101	5.387

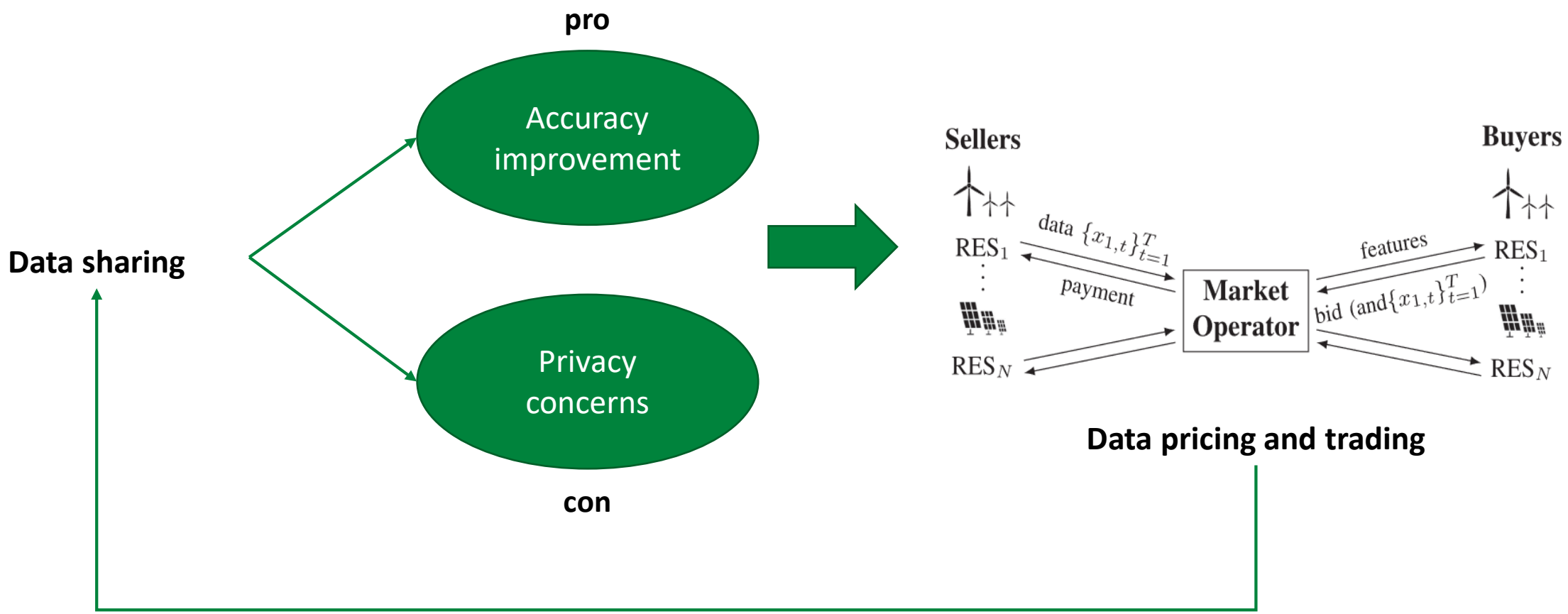


Performance improvement with different resolutions.

Data Sharing for probabilistic forecasting

Data Pricing in Forecasting

An incentive to accelerate data sharing



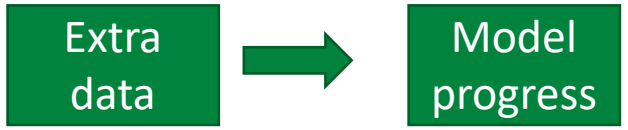
Data Sharing for probabilistic forecasting

Data Pricing in Forecasting

One-staged Pricing

One of the most common pricing strategy is pricing data based on its contribution on forecasting tasks. Thus, the definition of contribution tend to be evidently model-related:

e.g.
Reduction on loss function[1]
Improvement on forecasting accuracy
……



It can be called One-staged pricing due to pricing DIRECTLY based on the performance of additional data on the forecasting model.

[1] Pinson P, Han L, Kazempour J. Regression markets and application to energy forecasting[J]. arXiv preprint arXiv:2110.03633, 2021.

Data Sharing for probabilistic forecasting

Data Pricing in Forecasting

Two-staged Pricing

Beyond modelling directly based on regression models, pricing strategy can be two-staged, i.e. modelling the FURTHER revenue.

e.g.
Reduction on imbalance costs.[1]
Decrease on operational costs.[2]
……

Two-staged pricing tend to combine the techniques of forecasting and optimization, with respect to the knowledge of ML and economics.

[1]C. Gonçalves, P. Pinson, and R. J. Bessa, “Towards Data Markets in Renewable Energy Forecasting,” IEEE Transactions on Sustainable Energy, vol. 12, no. 1, pp. 533–542, Jan. 2021, doi: 10.1109/TSTE.2020.3009615.
[2]M. Yu et al., “Pricing Information in Smart Grids: A Quality-Based Data Valuation Paradigm,” IEEE Transactions on Smart Grid, Apr. 2022.

[1]

Data Sharing for probabilistic forecasting

Data Pricing in Forecasting

Extra data → Model progress → Lower costs

- RES agents' profit in energy market:

$$\rho(\hat{x}_t, x_t) = \pi_t^s x_t - C_t^{\uparrow/\downarrow}$$

where

$$C_t^{\uparrow/\downarrow} = \begin{cases} \lambda^{\uparrow}(\hat{x}_t - x_t), & \hat{x}_t > x_t \\ -\lambda^{\downarrow}(\hat{x}_t - x_t), & \hat{x}_t < x_t, \end{cases}$$

$$\lambda_t^{\uparrow} = \max(0, \pi_t^{\uparrow} - \pi_t^s),$$

$$\lambda_t^{\downarrow} = \max(0, \pi_t^s - \pi_t^{\downarrow}),$$

- RES agents' profit in data market:

$$\mathcal{G}_i(x_{i,t}; \mathbf{X}^{\mathrm{S}}, \mathcal{M}_i) = \left(\rho(\hat{x}_{i,t}^{\mathrm{market}}, x_{i,t}) - \rho(\hat{x}_{i,t}^{\mathrm{local}}, x_{i,t})\right)^+$$

Minimize $\mathcal{G}_i(x_{i,t}; \mathbf{X}^{\mathrm{S}}, \mathcal{M}_i)$

[1]C. Gonçalves, P. Pinson, and R. J. Bessa, "Towards Data Markets in Renewable Energy Forecasting," IEEE Transactions on Sustainable Energy, vol. 12, no. 1, pp. 533–542, Jan. 2021.

Data Sharing for probabilistic forecasting

Data Pricing in Forecasting

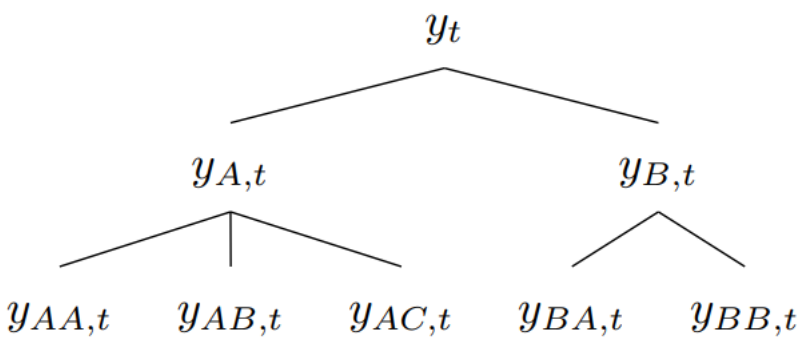
Fairly allocation: Shapley Value

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

In the context of forecasting, we can dim the total improvement on the forecasting task as the GOAL of the cooperative game and allocate the data buyers' payment among data sellers according to their marginal contribution.

Data Sharing for probabilistic forecasting

Hierarchical probabilistic forecasting



Example of a hierarchical time series .

1. Bottom-Up Probabilistic Forecasting
2. Mean Forecast Combination and Reconciliation

$$\mathbb{E}[\boldsymbol{y}_{T+h}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T] = \boldsymbol{S}\,\mathbb{E}[\boldsymbol{b}_{T+h}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T]$$

Taieb S B, Taylor J W, Hyndman R J. Coherent probabilistic forecasts for hierarchical time series[C]//International Conference on Machine Learning. PMLR, 2017: 3348-3357.

Data Sharing for probabilistic forecasting

Summary：

- Integrating various data;
- Protect privacy: encryption, DP, etc.;
- Federated learning: vertical and horizontal;
- Valuation/contribution of data.

Q&A

Part 3: State-of-the-art methods from statistical and machine learning

Instructor: Dr Yi Wang

The University of Hong Kong