Diffusion Models for Multivariate Probabilistic Net Load Forecasting in Transmission Grid Congestion Management

ENBIS, Florence - 9 September 2026

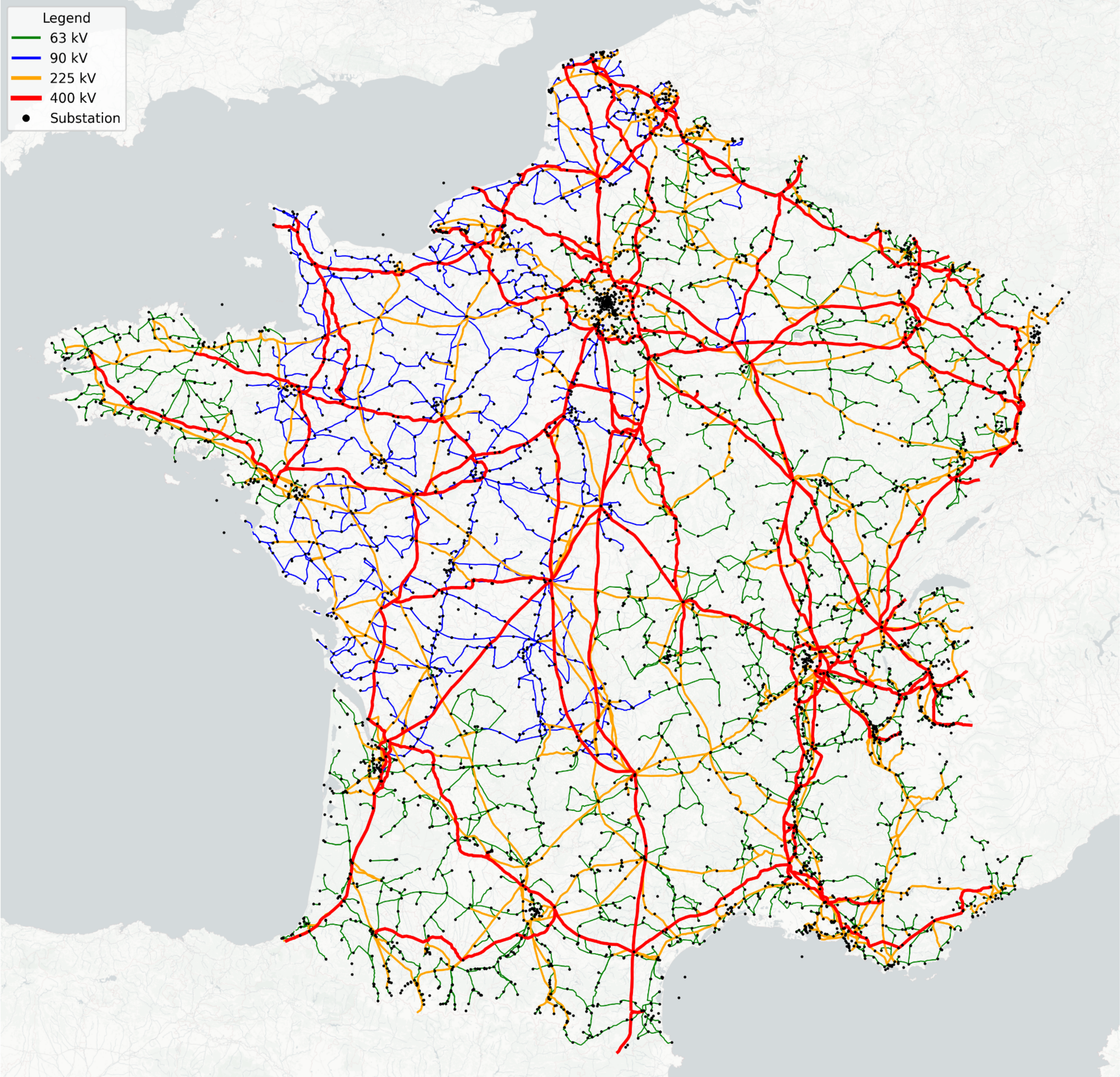
Maxime Ducourau, Ouassim Feliachi

Rte Le réseau de transport d'électricité

EPFL

RTE, French Transmission System Operator (TSO)

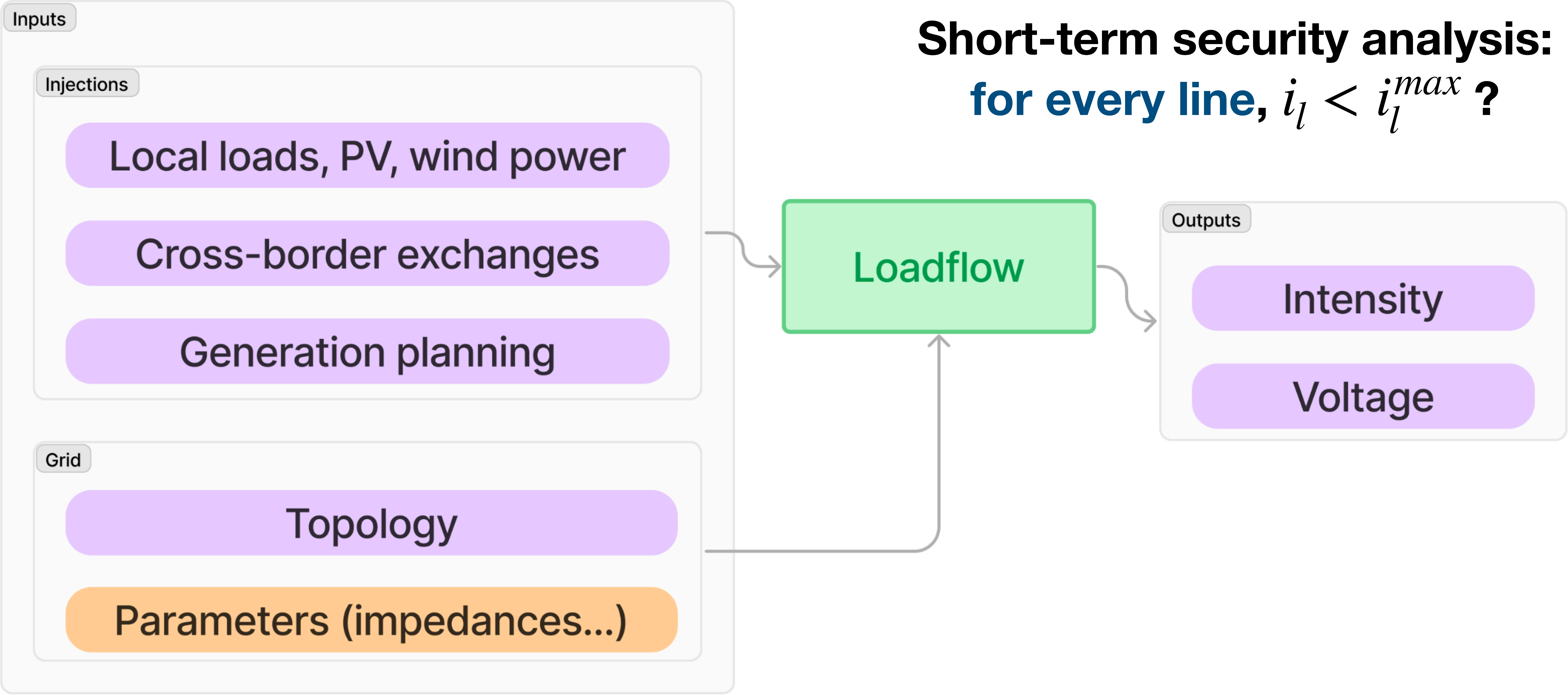
RTE maintains and develops the high and very-high voltage network.



Europe's largest transmission grid (105,000 km), across 4 voltage levels

→ Public-service obligation: guarantee electricity supply anytime, everywhere

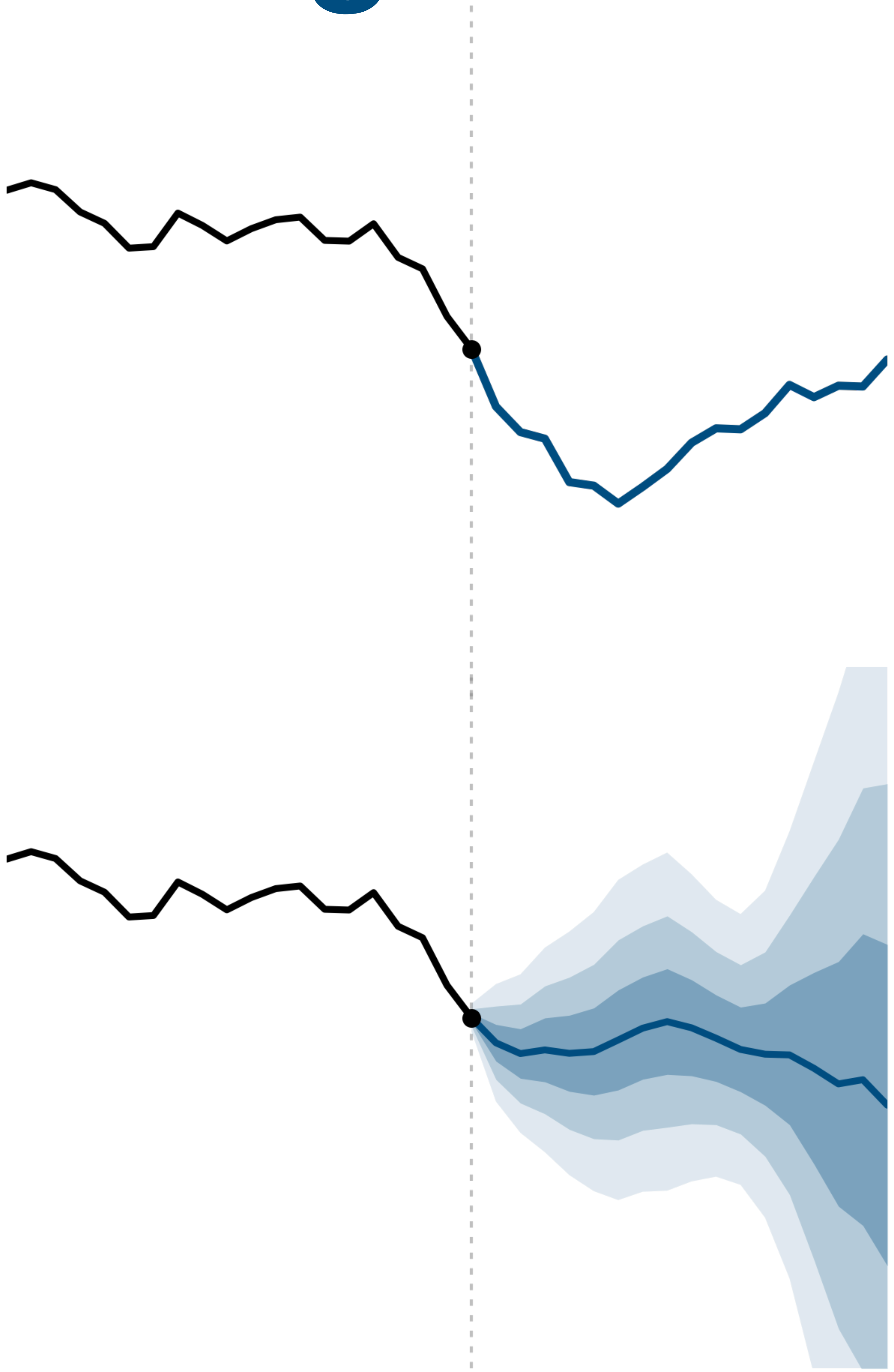
Operational need: security analysis



Short-term security analysis:
for every line, $i_l < i_l^{max}$?

With rising renewable penetration, inputs get more uncertain, and there is no uncertainty quantification

The case for probabilistic forecasting



Growing uncertainty calls for a probabilistic approach:

- **Obtain a measure of the uncertainty of the point forecast to give context to the dispatcher**
- **Logical extension of dispatcher behavior: testing different scenarios (e.g. bumping a PV installation output by 1 MW and observing the effect)**

→ **Required input to any stochastic optimisation of remedial actions**

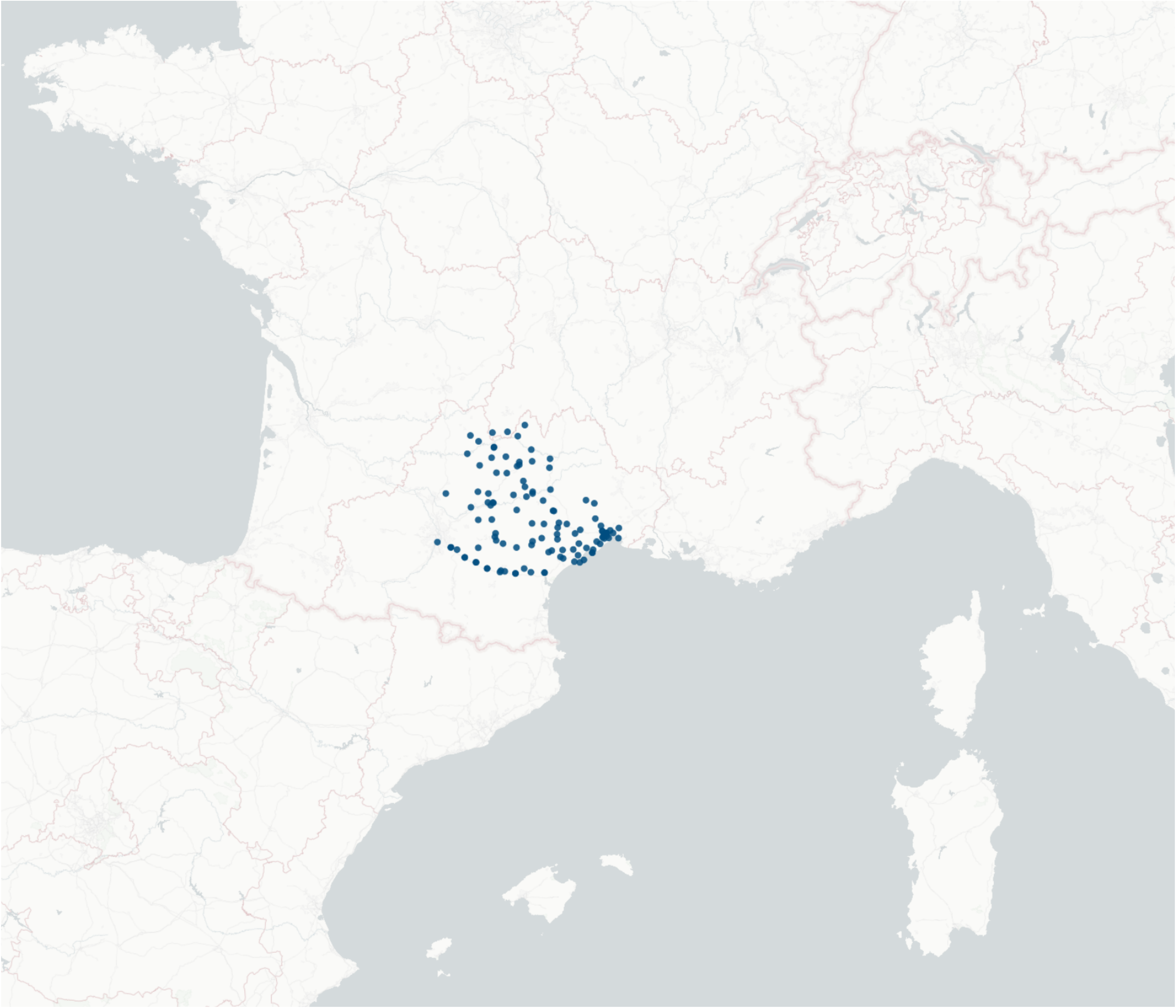
Data

We want to assess renewable production and load uncertainty. At RTE, only net load is telemetered, but it's enough for a load flow.

For a substation d : $\tilde{L}_d(t) = L_d(t) - G_d^{\mathrm{PV}}(t) - G_d^{\mathrm{wind}}(t)$

positive = withdrawal from grid

negative = injection back



Location of our 115 substations

115 substations in Toulouse area, 2017 - 2025

Profiles differ in kind: industrial, residential, PV-driven, wind-driven

Weather moves them together

Net load is volatile, heterogeneous and spatially correlated

Objective: joint distribution

Target: joint law of net load across substations $\in \mathbb{R}^D, D = 115, H = 3$

$$\mathbf{y}_{t+H}^{(k)} \sim p_\theta\left(\mathbf{y}_{t+H} \;\middle|\; \mathbf{c}\right)$$

$\mathbf{c}$: conditioning variables (weather, previous observations, calendar features)

A line flow is a (non-linear) function of injections at the substations. Thus, the marginals of the injections are not sufficient to obtain the marginals of the line flows.

→ **Well-calibrated scenarios at the substation level**

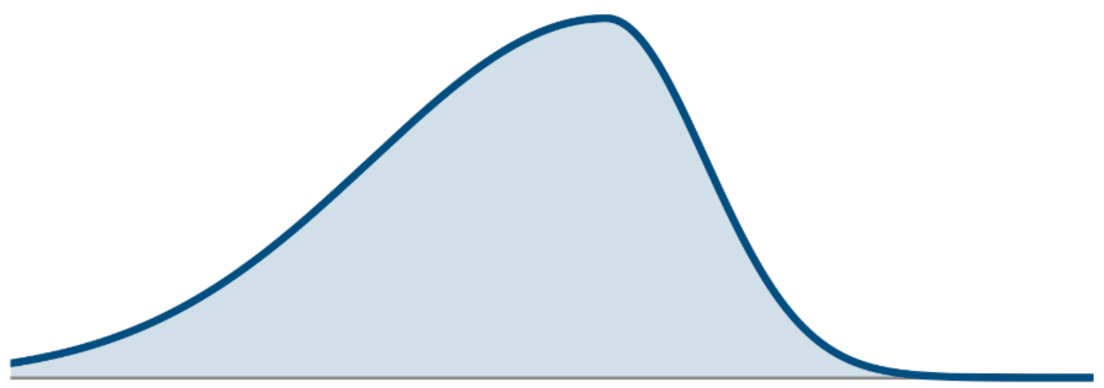
→ **Good reproduction of correlation structure between substations**

Three validation levels

$\mathbb{E}[y \mid c]$

1. **Deterministic:** « best forecast »

$p(y_d \mid c)$



2. **Probabilistic:** marginals

$p(y_1, \ldots, y_D \mid c)$

3. **Multivariate:** joint distribution

Related Work

Usually, we aim to build good marginals, then reconstruct the dependence

- **Marginals with quantile regression: linear models, GB, qGAM, NN approaches…**
- **Dependence structure (copulas): parametric or empirical (Schaake Shuffle [1])**

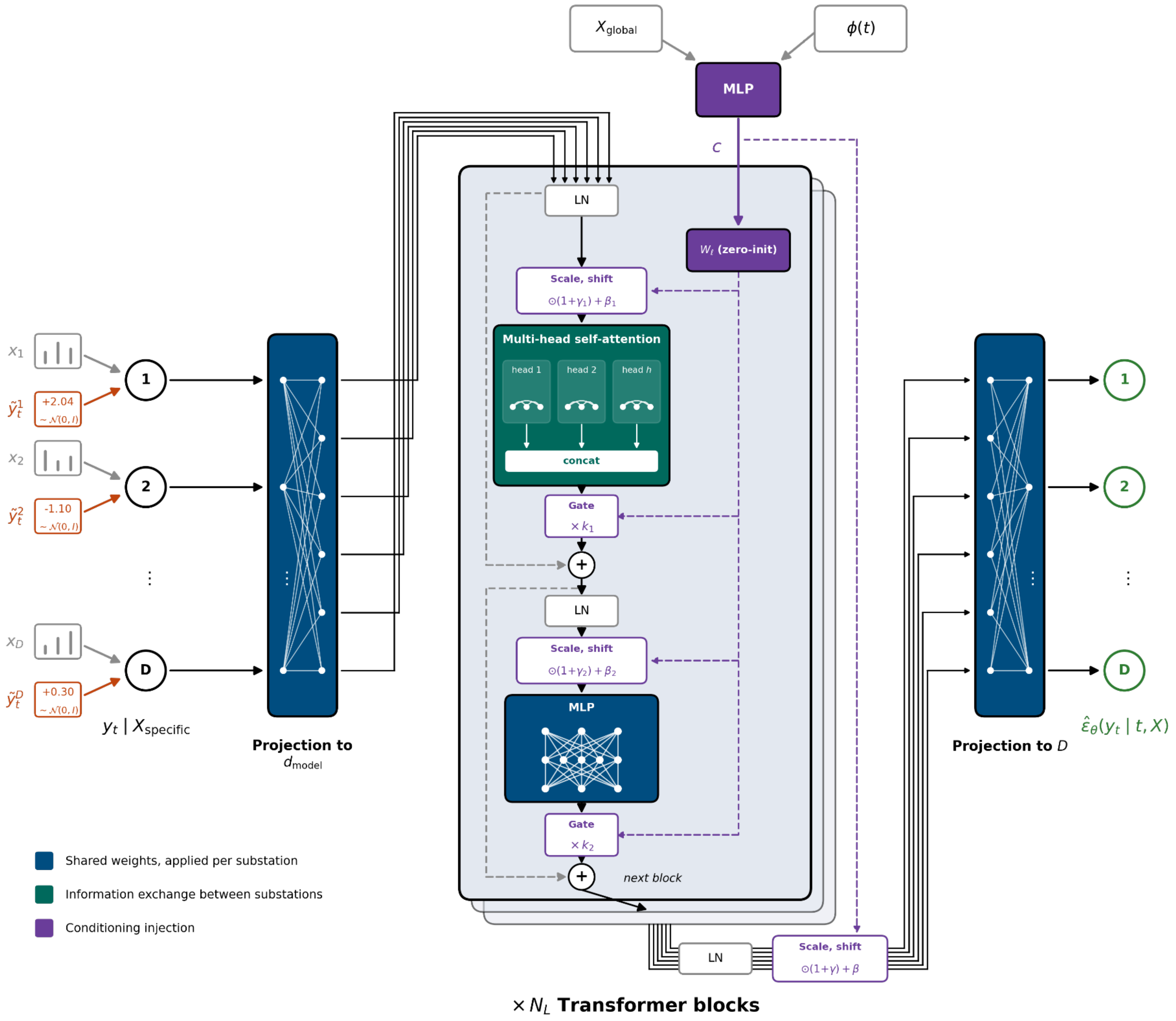
Diffusion for time series: large focus on temporal axis [2, 3]. Diffusion in energy: one site at a time [4]. Metrics don't focus on spatial dependence.

Generative model of the joint distribution (spatial correlation) and its validation protocol

Diffusion model

Diffusion models turn random noise into realistic samples from a target distribution. We learn to transport from one to the other.

- **Start from pure noise, denoise step by step until we get a plausible scenario**
- **At each step, the model estimates the noise to remove, guided by known info (weather, calendar…)**
- **Each substation = one token (its features + current noisy value)**
- **Substations learn their dependence structure using attention**



Baselines

Multivariate comparisons are built as marginals x copula:

- **Marginals: operational forecast + empirical CDF of residuals**
- **Marginals: Quantile Gradient Boosting (pinball loss)**
- **Dependence with Schaake Shuffle [1]: reordering based on historical ranks**

2 variants each: raw and recalibrated afterwards on validation set (one spread factor per substation, then one inflation of the mode shared by all)

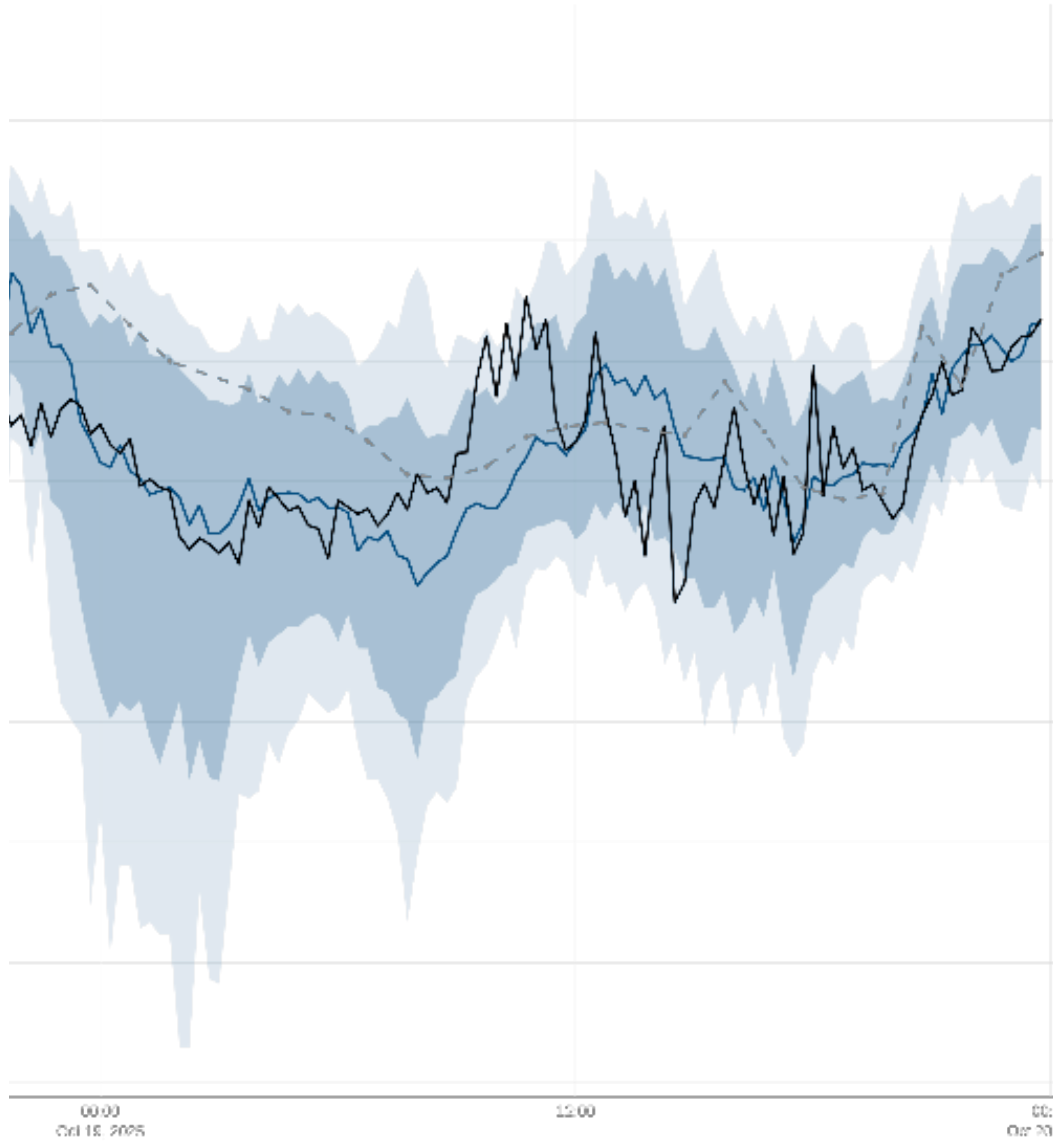
Diagnostics

1. Deterministic: MAE, RMSE (in skill score)
2. Marginal: CRPS (in skill score), coverage and pinball loss, per substation
3. Multivariate: Energy Score, Variogram Score [6]

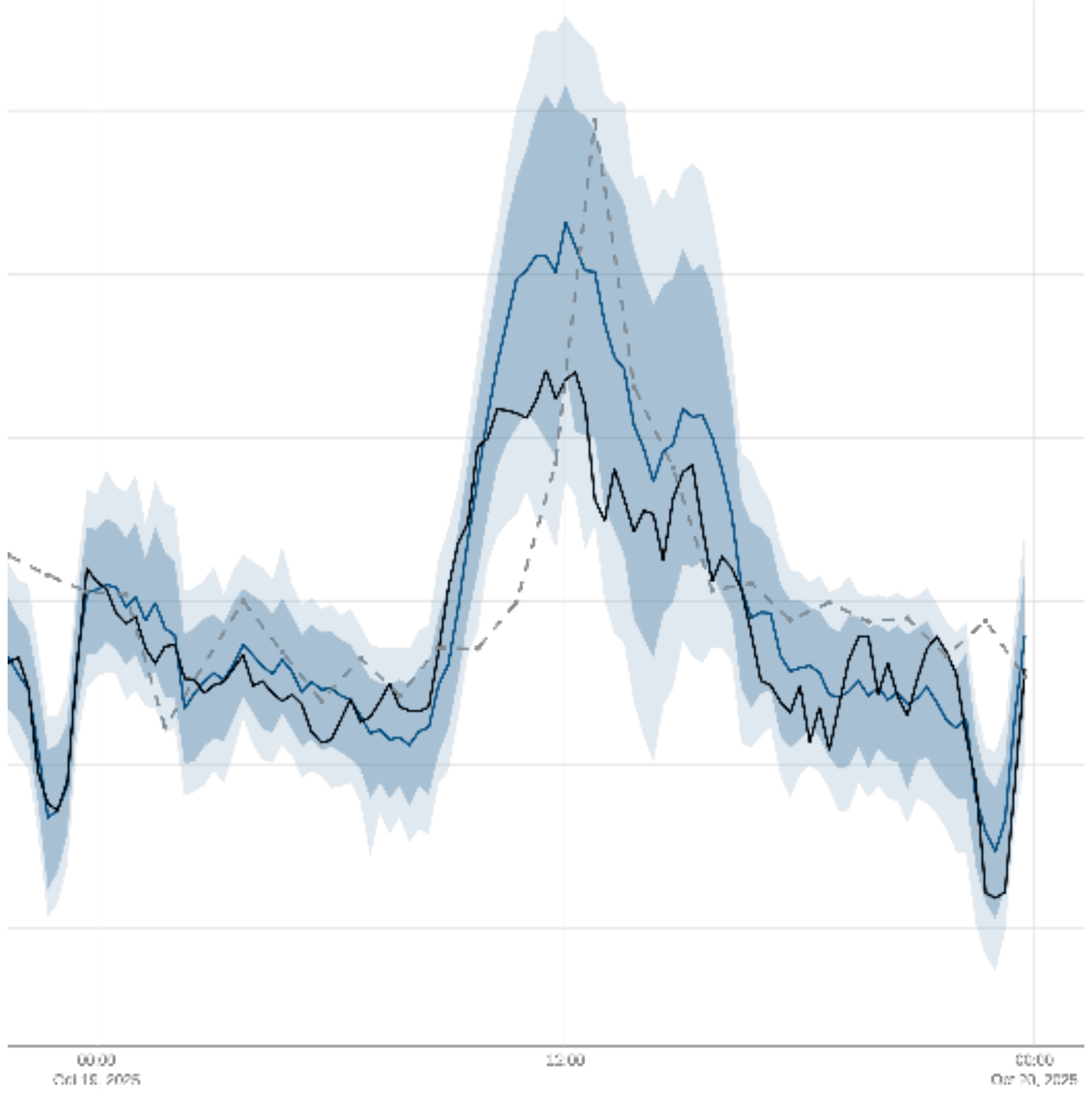
Problem: ES discrimination degrades with dimension [7]. VS dilutes pairwise sensitivity over D^2 pairs, thus we also score two projections:

- Aggregate net load (sum of all substations)
- Projected flows via PTDF (linear approximation of injections → flows)

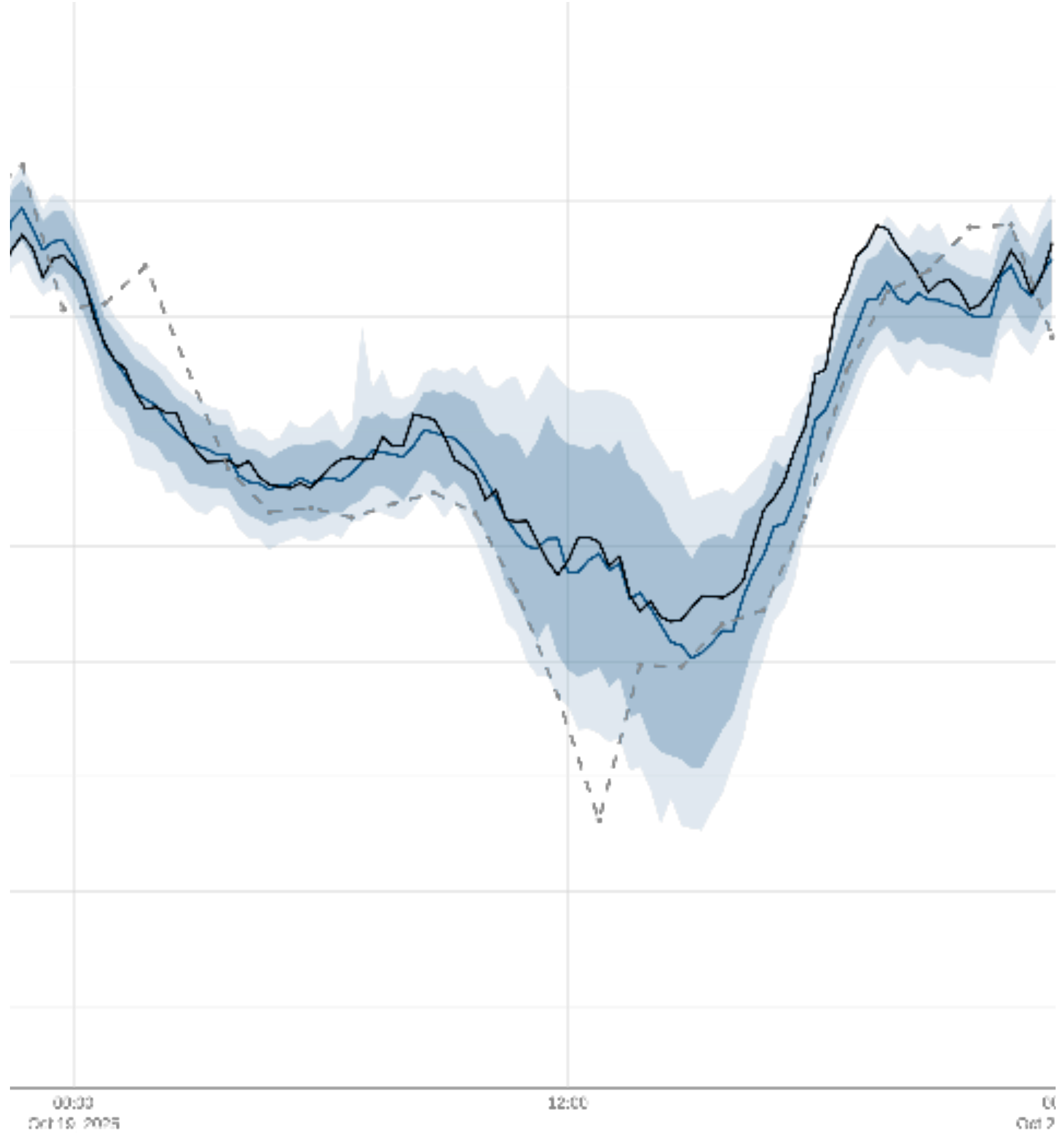
Forecast examples



Substation



Line flow



Aggregate

Diffusion model forecasts for a substation, a line flow and the aggregate, for October 19

Results: deterministic forecasts

Against persistence $y_{t+3} = y_t$ as the base of the skill score:

Model	SS_{RMSE} (%)	SS_{MAE} (%)
Persistence	0.0	0.0
Current operational model	9.3	9.8
Operational model dressing	11.0	11.0
Quantile GB	39.8	41.0
Diffusion model	**44.1**	**49.4**

$$\text{SS} = \left(1 - \frac{\text{Score}_{\text{model}}}{\text{Score}_{\text{reference}}}\right) \times 100$$

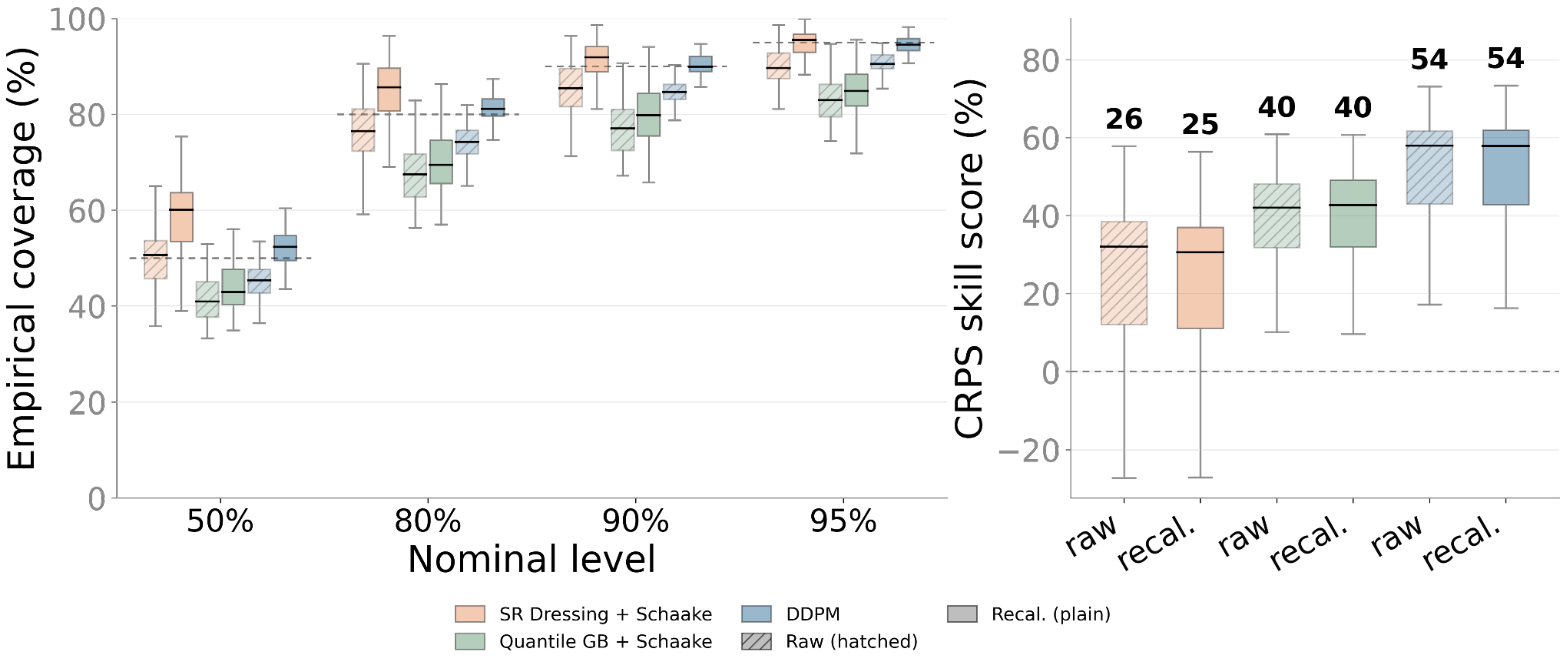
Results: marginals

Coverage of the prediction intervals and CRPS against dressed* persistence + Schaake Shuffle:

	Coverage (%)				CRPS (MW)		
Model	50%	80%	90%	95%	median	mean	SS_{CRPS} (%)
Dressed persistence + Schaake	46.5	73.9	85.2	90.0	2.30	2.31	0.0
Operational model dressing + Schaake	**54.7**	84.8	91.9	96.0	1.80	1.91	9.2
Quantile GB + Schaake	51.0	**79.1**	87.7	89.2	1.12	1.31	42.9
Diffusion model	52.6	81.5	**90.3**	**94.9**	**0.85**	**1.08**	**52.4**

**attaching error samples to build an ensemble*

Results: joint distribution, flows



Is it only due to marginals' quality?

Took DDPM marginals, and switched dependence structures:

Metric	Diffusion model + Indep.	Diffusion model + SS	Full diffusion model
Energy Score	20.409	20.356 (0.3 %)	**20.280** (0.6 %)
Variogram Score	0.2958	0.2949 (0.3 %)	**0.2946** (0.4 %)
CRPS substations (mean)	1.0807	1.0807 (0.0 %)	1.0807 (0.0 %)
CRPS flows (mean)	0.8958	0.8901 (0.6 %)	**0.8827** (1.5 %)
CRPS flows (median)	0.4567	0.4397 (3.7 %)	**0.4123** (9.7 %)
CRPS_{sum}	49.609	46.840 (5.6 %)	**43.928** (11.5 %)
$\text{QL}_{10}^{\text{sum}}$	19.682	15.981 (18.8 %)	**13.484** (31.5 %)

→ **No, and the standard multivariate scores can't see the difference!**

Takeaways

1. **Generating the joint distribution beats assembling it** and the gain lands on the aggregated quantities, not on the marginals
2. **The standard multivariate scores don't see it** → project onto the quantities decisions are taken on
3. The denoiser holds **no parameter per substation scaling**, and forecasting unseen substations, are changes of scale, not of model

Next steps

Scaling up to France (≈ 2,500 substations): hierarchically or not. With no additional effort, the model already runs on the Toulouse region (511 substations).

Multi-step-ahead forecasting.

Probabilistic forecasting of NAZA activation: support the decision to end curtailment, using a curtailment probability output by the model (historical curtailment data is available, and in zones where flow uncertainty is mainly driven by renewables).

After scaling to France → scenario clustering → generation of probabilistic variants to guide the operator's complementary studies, time permitting.

References

[1] *Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B. & Wilby, R. (2004). The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields, Journal of Hydrometeorology*

[2] *Rasul, K., Seward, C., Schuster, I. & Vollgraf, R. (2021). Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting. ICML*

[3] *Tashiro, Y., Song, J., Song, Y. & Ermon, S. (2021). CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation, NeurIPS*

[4] *Dumas, J., Wehenkel, A., Lanaspeze, D., Cornélusse, B. & Sutera, A. (2023). Denoising Diffusion Probabilistic Models for Probabilistic Energy Forecasting, IEEE Belgrade PowerTech*

[5] *Ho, J., Jain, A. & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models, NeurIPS*

[6] *Scheuerer, M. & Hamill, T. M. (2015). Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities, Monthly Weather Review*

[7] *Pinson, P. & Tastu, J. (2013). Discrimination Ability of the Energy Score, DTU Compute Technical Report*

Questions?

Appendix: recalibration

Two-stage correction, fit on a small calibration split, applied to test.

Aggregate scores can hide local miscalibration.

1. Marginal (per substation): scalar λ_d rescales each substation's residual around its median, it preserves rank.

2. Aggregate (shared residual): scalar η inflates/shrinks the cross-substation residual, fixing the sum's dispersion and leaving marginals untouched.

3. Both fit once on a small calibration split (1000 draws), applied as-is to test.

Appendix: dressing

Add K historical persistence errors, sampled empirically at each substation, to turn the point forecast into an ensemble.

Each ensemble member adds one historical error to the point forecast:

$$\hat{y}_{n,j}^{(k)} = \hat{y}_{n,j} + \varepsilon_j^{(k)}, \quad k = 1,\ldots,K$$

The errors are drawn empirically from past persistence residuals at that substation:

$$\varepsilon_j^{(k)} \sim \{y_{m,j} - \hat{y}_{m,j} : m \in \text{calibration window}\}$$