ICHQP 2026
22nd International Conference on Harmonics and Quality of Power
Dresden, Germany

#24

Ensemble Forecasting of Power Quality Parameters

Max Domagk, Peter Feistel, Jan Meyer, Marco Lindner, Jako Kilter

Motivation

Current developments

- Increase in renewables and new technologies (e.g. HVDC) → more power electronics
- Potential impact on PQ levels (e.g. harmonics) → extended PQ monitoring campaigns
- Large amounts of data (often only compliance assessment) → underutilized potential

Forecasting of PQ parameters

- Early detection of critical developments
- Proactive asset management
- Open challenges (data heterogeneity, seasonality, predictability)

Motivation

State of the art

- Short-term horizons (hours to days), selected sites or parameters
- Strong focus on Machine Learning models
- Ensembles established in load forecasting, rarely used for PQ

Gap

- Reliable medium- and long-term forecasts across many PQ parameters and sites

This paper

- Systematic evaluation of individual and ensemble forecasts
- 716 weekly time series, German and Estonian transmission system

PQ Parameters

Voltage quality

- Continuous PQ parameters; not PQ events (e.g. dips, swells, transients)
- Aggregated measurement values (10-min values acc. to IEC 61000-4-30)
- Parameters:
  - Long-term flicker (Uplt)
  - Voltage unbalance (UNB)
  - Total harmonic distortion (Uthd)
  - Individual harmonics (U03 … U15, odd)

Characteristics

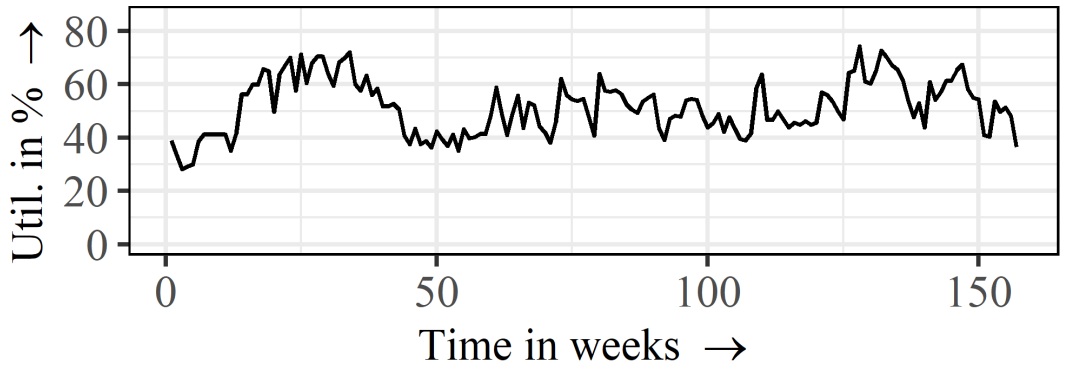
- Site- and parameter-dependent (e.g. flicker behaves differently than harmonics)
- Time-dependent (trends and seasonality vary strongly, partly irregular)

Measurement Data

Dataset

- Measurement campaigns in Germany and Estonia
- 27 sites in transmission system (110–380 kV), at least 3 years
- 716 time series in total

Utilization of unbalance (DE, 220 kV)



Preprocessing

- 10-min values → weekly 95th percentiles
- Filling missing weeks (max. 20 %, filled with last available value within 10 weeks)
- Normalizing by planning level limits (VDE-AR-N 4120/4130, IEC TR 61000-3-6)

→ Utilization

Forecasting Models

Selection criteria

- Simple, robust, no per-time-series tuning (scalable to hundreds of time series)
- Automatic model parameter selection

Individual models

- Benchmark: Seasonal Naive
- Direct modelling: Holt-Winters, SARIMA, Prophet
- Decomposition + forecast: STL-Drift, STL-ES, STL-Holt, STL-ARIMA

Forecasting:

- Training: first 105 weeks → forecast horizon: 52 weeks (1 year)
- sMAPE as error metric (percentage error, between 0-200 %)

Ensemble Framework

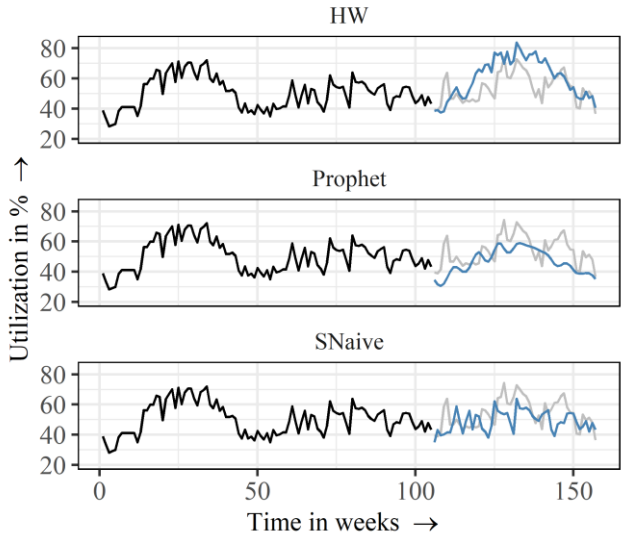
- No single model is best for all time series → combine instead of select

Individual forecasts

8 individual models
(SNaive, HW, SARIMA, Prophet, STL-*x*)

Ensemble Framework

- No single model is best for all time series → combine instead of select



Individual forecasts

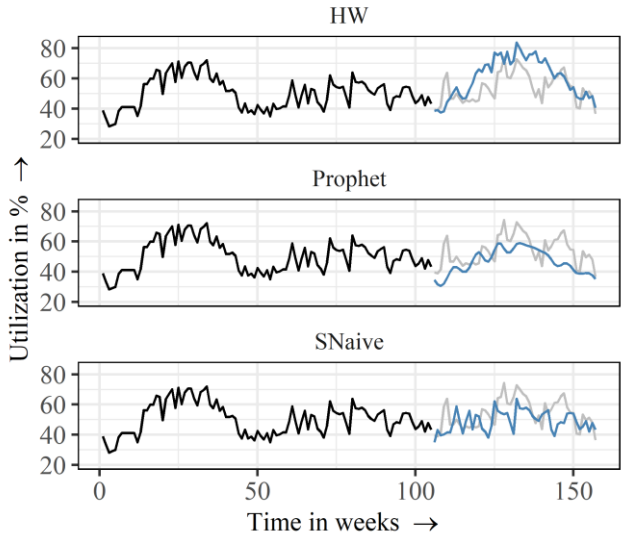
8 individual models
(SNaive, HW, SARIMA, Prophet, STL-*x*)

Model selection

Which models enter the ensemble
(all combinations of ≥2 models → 247)

Ensemble Framework

- No single model is best for all time series → combine instead of select



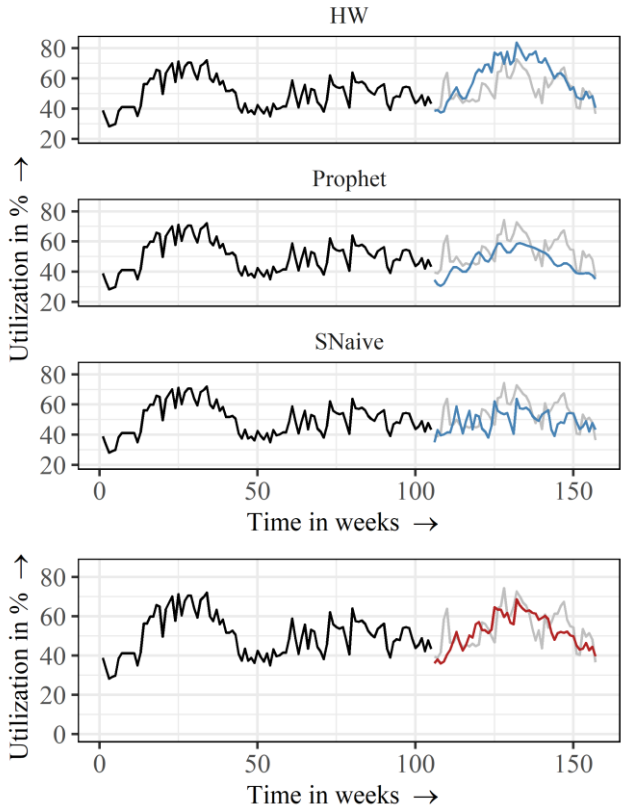
Individual forecasts
8 individual models
(SNaive, HW, SARIMA, Prophet, STL-*x*)

Model selection
Which models enter the ensemble
(all combinations of ≥2 models → 247)

Forecast combination
How forecasts are combined
(mean, median, weighted by sMAPE/rank)

Ensemble Framework

- No single model is best for all time series → combine instead of select



Individual forecasts
8 individual models
(SNaive, HW, SARIMA, Prophet, STL-*x*)

Model selection
Which models enter the ensemble
(all combinations of ≥2 models → 247)

Forecast combination
How forecasts are combined
(mean, median, weighted by sMAPE/rank)

Ensemble forecast
988 ensemble models per time series
(247 combinations x 4 combination methods)

Performance Evaluation

Metrics

- sMAPE over horizon h = 52 weeks
- Classification: good (sMAPE < 10 %), acceptable (sMAPE < 25 %), poor above

Procedure

1. Accuracy over full horizon (sMAPE) per model and time series
2. Model rank R(sMAPE) per time series
3. Mean accuracy ($\overline{\mathrm{sMAPE}}$) and mean rank ($\overline{\mathrm{R}}$) across all 716 series

Benchmark ratio $$\mathrm{BR} = \frac{\overline{\mathrm{sMAPE}}_{\mathrm{Model}}}{\overline{\mathrm{sMAPE}}_{\mathrm{SNaive}}}$$ (values < 1 indicate improvement)

Results – Individual Models

Decomposition-based models dominate

- STL-ARIMA best individual model (sMAPE = 18.22 %, BR = 0.873)
- STL-ES close second (18.67 %)
- SARIMA third (19.38 %)

Only 3 of 8 models beat the SNaive benchmark (20.88 %)

- Prophet, STL-Holt, STL-Drift and HW fall behind (up to 25.03 %)
- More complex ≠ better

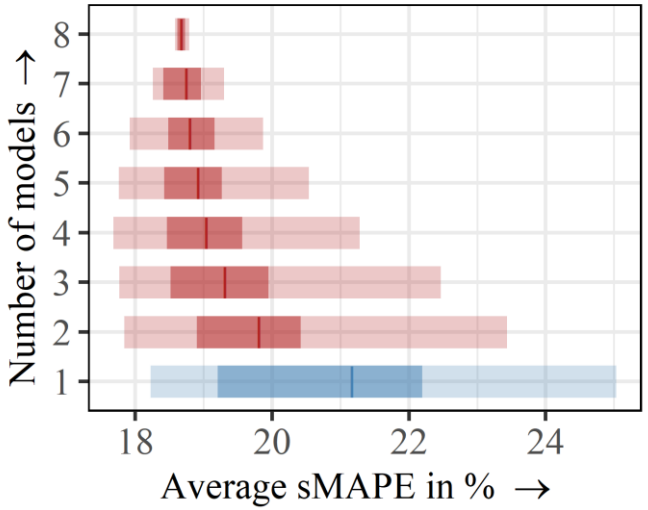
Results – Ensemble Models

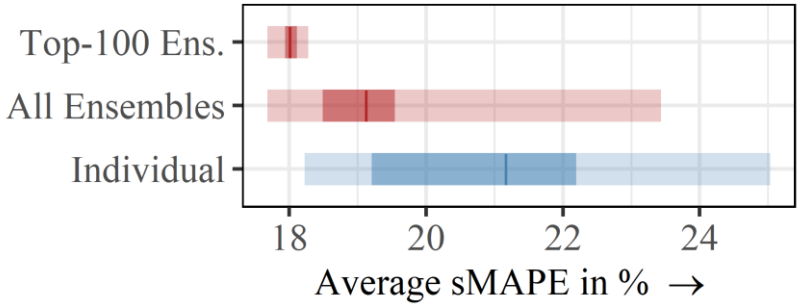
Best ensemble

- D28 Median (STL-ARIMA, STL-ES, SARIMA, SNaive) (sMAPE = 17.68 %, BR = 0.847)
- 15.3 % better than SNaive, 2.6 % better than best individual model
- Best individual model not within top 100 ensembles

Ensemble performance

- More models → lower error and lower variance
- Top 100: mostly 4-5 models from the best individuals
- Aggregation method secondary (composition matters more)





Conclusion

Findings

- No model is universally optimal (decomposition-based models perform best individually)
- Ensembles reduce model-selection risk (better accuracy, fewer extreme forecast errors)
- Composition matters more than aggregation method (mean or median is enough)
- Ensembles are scalable and computationally cheap
  (applicable to hundreds of measurement sites)

Outlook

- Probabilistic forecasting, asymmetric accuracy measures near PQ limits
- Open question: which time series are predictable at all?

Thank you for your attention!

Technische Universität Dresden

max.domagk@tu-dresden.de
+49 351 463 35223
maxdomagk.de