Neural Posterior Estimation for Empirical Power System Time Series

Ulrich Oberhofer
Karlsruhe Institute of Technology
Karlsruhe, Germany
ulrich.oberhofer@kit.edu

Nicolas Weber
Karlsruhe Institute of Technology
Karlsruhe, Germany
nicolas.weber@student.kit.edu

Veit Hagenmeyer
Karlsruhe Institute of Technology
Karlsruhe, Germany
veit.hagenmeyer@kit.edu

Benjamin Schäfer
Karlsruhe Institute of Technology
Karlsruhe, Germany
benjamin.schaefer@kit.edu

Abstract

Methods for state and parameter estimation are widely used to analyze complex systems, such as power systems. Estimating parameters is often necessary for follow-up simulations or effective control. Bayesian approaches allow us to go beyond point estimates and include domain knowledge when identifying parameters. However, such Bayesian approaches are often limited to simulated settings and not well-tuned for noisy, empirical data. We propose to investigate power systems, specifically the power grid frequency dynamics, as an important empirical use case with non-trivial data. Within our article, we introduce the concept of Simulation-Based Inference (SBI) for power grid frequency dynamics. Furthermore, we identify an amortized Neural Posterior Estimator (NPE), using recent normalizing flow architectures, as a suitable method for Bayesian parameter inference. We initially validate the model based on simulated observations and demonstrate the usability of amortized NPE on stochastic models for power grid frequency. Next, we apply the estimation method on empirical frequency data from European synchronous areas and quantify how it outperforms three baseline models based on newly designed posterior predictive checks. Finally, we carry out a detailed analysis of the inferred posterior distributions, in terms of time-dependency and correlations, providing some additional consistency checks. Overall, we demonstrate how Bayesian inference can be carried out and quantified in a complex, noisy, empirical setting.

CCS Concepts

• **Applied computing → Mathematics and statistics**; **Engineering**; • **Mathematics of computing → Mathematical analysis**; • **Hardware → Power and energy**; • **Computing methodologies** → *Machine learning*; *Artificial intelligence*.

Keywords

Simulation-based Inference, Neural Posterior Estimation, Power Grid Frequency Dynamics, Stochastic Differential Equations, Ill-posed Problems, Inverse Problems

ACM ISBN 979-8-4007-2011-6/26/06
https://doi.org/10.1145/3744255.3811710

ACM Reference Format:
Ulrich Oberhofer, Nicolas Weber, Veit Hagenmeyer, and Benjamin Schäfer. 2026. Neural Posterior Estimation for Empirical Power System Time Series. In *The 17th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3744255.3811710

1 Introduction

State and parameter estimation are fundamental methods for understanding complex systems. Inferring uncertain quantities from empirical observations allows us to draw substantial conclusions about underlying system behavior, which, in turn, allows us to generate forecasts or efficient system control [21]. Bayesian inference is a widely used method that supports the estimation of uncertain quantities in distributions. It involves the update of existing belief (prior) after the observation of new data, which results in the posterior estimate. Bayesian methods have been heavily involved in parameter estimation research [15]. In many applications, Bayesian estimation faces the challenge that the probability distribution of observations is not tractable. One possible solution is offered by Approximate Bayesian Computation (ABC) [34, 40] or Variational Inference [2]. Methods of simulation-based inference (SBI), as a subfield of Bayesian inference, are increasingly being applied in various areas of natural science. In this field, the complexity of estimation methods has grown significantly in recent years. Sampling models as ABC rejection [1, 34] apply repeated sampling from the prior distribution, generating a simulation via the underlying model, and add samples to the posterior distribution according to a selected distance to the ground-truth observation.

Recently, methods in SBI have increased in complexity. Powerful machine-learning based models such as Neural Posterior Estimation (NPE) [23, 30], Neural Likelihood Estimation [22, 32], and Neural Ratio Estimation [8, 13] have introduced new, amortized methods for inferring posterior parameter distributions from observed data with deep neural networks, built with different architectures of neural density estimation [6]. In practice, SBI often faces challenges because real-world data rarely align perfectly with simulator assumptions, leading to model-data mismatches that can impair posterior inference [14, 45].

Power grid frequency is a crucial indicator for the stability of power networks. The increasing share of renewable energy sources in the grid poses new challenges for ensuring stability [37]. The

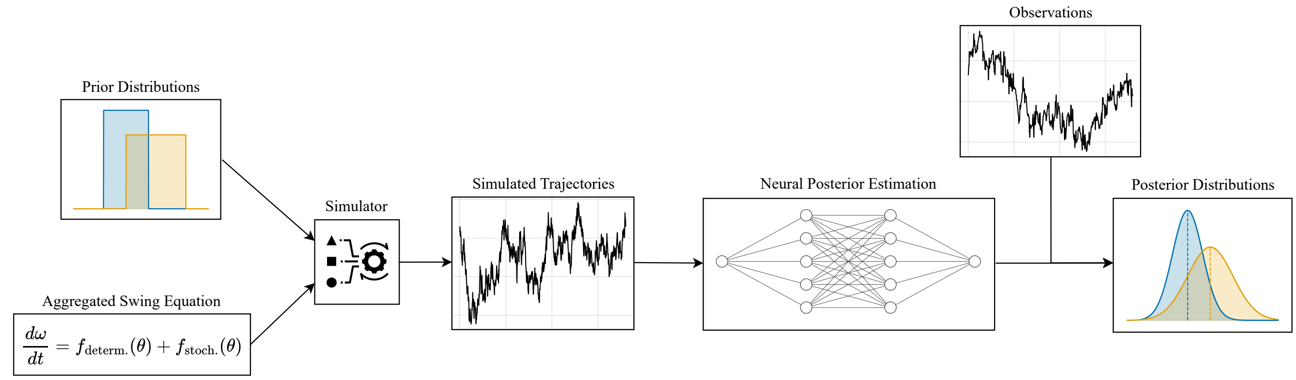

Figure 1: Schematic representation of our model approach of neural posterior estimation for the time-specific determination of parameters of empirical power grid frequency time series. With the use of a stochastic differential equation (SDE) and prior assumptions of the equation parameters, time series segments are simulated. A neural density estimator in the form of a normalizing flow is trained to approximate the posterior distribution based on simulated observations. The amortized posterior estimation method then enables inference of the posterior distribution based on new observations. For each time series segment, a posterior distribution over the SDE parameters is inferred.

guarantee of security of power systems is based on reliable parameter and state estimation [49]. Thus, assessing central variables in power systems, such as the primary control strength representing the ability of a power system to return to a stable state, is important. In a similar spirit, secondary control, power imbalances, and ambient frequency noise all have a detrimental impact on the stability of a power system if they are out of proportion in relation to the control mechanisms. Here, estimating both the parameters and their uncertainty can serve an important role for control, as large distributions of a particular fundamental parameter can lead to large frequency fluctuations or excursion, or, in the worst-case scenario, partial or total blackouts.

Classical state and parameter estimation methods in power systems [20] often rely on low-dimensional parametric models, simplified noise assumptions, and iterative optimization tailored to steady-state or quasi-static regimes. However, frequency measurements in modern power grids are high-dimensional, strongly non-Gaussian, and non-stationary. They are also influenced by partially unknown and rapidly changing dynamics, such as inverter-based resources and stochastic disturbances. These properties render repeated likelihood evaluation and model re-specification impractical, thereby motivating the use of amortized neural density estimators that can learn complex conditional distributions directly from data and simulations. In the field of machine learning, various methods are used in different applications, ranging from time-series forecasting of load and renewable generation to state and parameter estimation for analyzing system dynamics.

Bayesian methods are used in a variety of applications of energy systems. Bayesian physics-informed neural networks have been shown to outperform conventional system identification methods in noise-prone data [38, 39]. For detecting large frequency deviations, Bayesian decision methods [19] and Bayesian structure frameworks [24] are applied.

The development and application of state and parameter estimation methods for power systems is a highly active research field both for synthetic experiments and for the analysis of empirical data [43, 48]. In the field of frequency stability analysis, the influence of techno-economic features on frequency indicators is explained using probabilistic methods in [27]. In [17], the information from electricity market features is extracted and combined with neural networks and analytical models to enable parameter inference to model power grid frequency. In [44], NPE, in combination with ABC, is used for parameter calibration in power systems. However, NPE or SBI in general have not yet been applied to empirical power grid data.

In the present paper, we introduce the application of neural posterior estimation on stochastic power system time series. To the best of our knowledge, this study represents the first application of SBI on empirical power system data.

Our main contributions are as follows:

(1) We apply neural posterior estimation on stochastic differential equations that represent the dynamics of power grid frequency, and validate the inferred posterior distributions on simulated target data.
(2) We demonstrate real-world applicability by successfully applying our method to empirical frequency data from two major European synchronous areas, establishing neural posterior estimation as a viable tool for parameter inference in operational power systems.
(3) We analyze the estimated parameter distributions on time-dependent patterns, and highlight differences in the observations between the Continental and the Northern European power systems.
(4) We analyze the relationship between the variance of frequency time series segments and the uncertainty regarding the estimated posterior parameter values. In both analyzed power grids, we observe a negative correlation between the variances of the trajectory and the estimated frequency control parameters, as well as a positive correlation between

the variances of the trajectory and the parameters which describe the power imbalance.

The following sections of the paper are structured as follows: In Section 2, we introduce the general structure of the neural posterior estimation and the physical model that represents power grid frequency dynamics, and in Section 3, we describe the experimental set-up of our analysis, as well as the validation and baseline methods. We quantitatively validate our approach in Section 4. In Section 5, we apply neural posterior estimation on empirical power grid frequency data from the Continental European (CE) synchronous area and the synchronous grid of Northern Europe. Here, particularly, we analyze the inferred posterior distribution with respect to time-dependent behavior and correlations, before concluding the present paper and providing insight into future work in Section 6.

2 Methodology

We first introduce the concept of NPE and how we obtain posterior distributions for a stochastic model. Next, we introduce a stochastic differential equation that describes power grid frequency dynamics and explain how we use it to generate synthetic trajectories for simulation-based inference.

2.1 Neural Posterior Inference

The general concept of Bayesian inference is the estimation of a posterior distribution via Bayes' theorem

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}, \tag{1}$$

where we refer to θ as a parameter and x as an observation. Bayesian inference aims to update the prior belief with respect to observed data. Here, $p(\theta \mid x)$ represents the posterior probability, $p(\theta)$ is the prior, $p(x \mid \theta)$ the likelihood, and $p(x)$ is the evidence with respect to the marginal property. Starting from a prior distribution, based e.g. on domain knowledge, the belief about θ is updated based on observation x, and the posterior probability is obtained.

In order to apply simulation-based inference, we first require a mechanistic model to generate simulations that aim to approximate empirical observations, and second, a prior distribution for the parameters of the mechanistic model. The prior distribution is based on physical constraints and domain knowledge, and serves as a basic assumption about the parameter distribution. We sample from a prior parameter distribution $p(\theta)$ and simulate realizations of the stochastic differential equation. These simulations are used in the following step to train a neural density estimator. Finally, we receive a posterior function $p(\theta \mid x)$ such that we can sample from the posterior conditioned on observation x.

To estimate the posterior probability distribution, we use masked autoregressive flows (MAFs) [31], a class of normalizing flows [33]. Normalizing flows are neural networks that transform a simple base distribution into a complex one using a sequence of invertible, differentiable mappings. They allow exact likelihood computation and flexible density estimation by tracking how probabilities change through these transformations. The masked autoregressive flow consists of several stacked masked autoregressive density estimators (MADE) [11].

While it does not rely on a likelihood calculation, the MAF estimates the posterior distribution $p(\theta \mid x)$ directly. The model learns a conditional density $q_\Theta(\theta \mid x)$, where Θ denotes the parameters of the MAF. This approximation can be understood as minimizing the Kullback-Leibler (KL) divergence between the true posterior and the learned distribution, averaged over the joint distribution $p(x, \theta)$. This yields the loss function

$$\mathcal{L}(\Theta) = -\mathbb{E}_{p(x,\theta)}\left[\log q_\Theta(\theta \mid x)\right]. \tag{2}$$

Once the posterior estimator has been trained, a new observation x_{obs} can be used to obtain samples from the approximate posterior $q_\Theta(\theta \mid x_{\text{obs}})$.

For our methodology and analysis, we use the implementation of neural posterior estimation (NPE-C) in [3].

2.2 Power-Grid Frequency Dynamics

Power grid frequency is a key indicator of a power system's stability. Frequency dynamics are influenced by various deterministic and stochastic effects. The main contributors are the balance between power generation and load, frequency control schemes, and noise on a small time scale. Aggregating the power grid frequency by averaging it over all nodes, we move into the center of inertia of the power grid [42] and obtain the stochastic differential equation [35]

$$M\frac{\mathrm{d}}{\mathrm{d}t}\omega = \Delta P_{\text{control}} + \Delta P_{\text{imbalance}} + \Delta P_{\text{fluct.}}. \tag{3}$$

Here, the dynamics of the angular velocity ω, which is connected to the grid's frequency as $\omega = 2\pi f - f_{\text{ref}}$ with the reference frequency $f_{\text{ref}} = 50$ or $60\ Hz$, depends on several power inequalities. Our model is based on previous research and modeling [17, 46].

$\Delta P_{\text{control}}$ represents the load-frequency control mechanisms in the system and is composed of primary and secondary control. While primary control acts on short time horizons of up to 30 seconds, secondary control has a higher reaction time of up to 15 minutes. Further, primary control serves to stabilize and attenuate frequency fluctuations in the short term, while secondary control aims to reduce the frequency deviation back to the reference frequency, and is often modeled as an integral controller. The control power imbalance is therefore approximated as

$$\Delta P_{\text{control}}(t) = M\left(-c_1\omega - c_2\int \omega(\tau)d\tau\right) = M\left(-c_1\omega - c_2\phi\right).$$

$\Delta P_{\text{imbalance}}$ describes the imbalance between load and the scheduled generation by dispatchable power plants. This schedule is regulated in blocks of $15, 30$, or 60 minutes, depending on the power system [47]. While energy consumption is a continuous process, deterministic power imbalances arise, which lead to deterministic frequency deviations (DFDs) [18, 46]. Due to the combination of a continuous and a steady process, these frequency imbalances can be modeled during a market schedule interval as a sawtooth function

$$\Delta P_{\text{imbalance}}(t) = M(P_0 + P_1 t).$$

Further, we aggregate all other unknown influences, e.g. from volatile renewable or quickly changing demand, into a fluctuation term, which we model as uncorrelated Gaussian noise:

$$\Delta P_{\text{fluct.}} = M\epsilon\xi(t),$$

where $\xi(t)$ represents white noise which is distributed according to the standard Gaussian distribution. Thereby, ϵ indicates the amplitude of fast imbalance fluctuations. The use of Gaussian-distributed noise is a simplification that is particularly applicable to large power grids [35].

To estimate unique parameter values rather than just ratios, we need to rescale the model parameters by the inertia constant M. Note that the stochastic differential equation stays invariant under the scaling. If we substitute the terms for $\Delta P_{\text{control}}$, $\Delta P_{\text{imbalance}}$, and $\Delta P_{\text{fluct.}}$ into (3), we obtain the stochastic differential equation

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\phi &= \omega \\ \frac{\mathrm{d}}{\mathrm{d}t}\omega &= -c_1\omega - c_2\phi + P_0 + P_1 t + \epsilon\xi(t). \end{aligned} \tag{4}$$

The physical model described in (4) serves to explain frequency dynamics and, under the assumption of suitable parameters, to simulate frequency trajectories. As frequency dynamics depend, as described, on multiple influences such as the composition of the electricity mix and the current load, the model parameters must also vary in order to model the frequency dynamics appropriately. Hence, we apply posterior estimation on time series segments of 15 minutes, according to the dispatch schedule of conventional power plants, focusing on both the Continental European and Nordic synchronous areas.

With our NPE approach, we are substantially expanding prior work with this stochastic model. While it is possible to derive the likelihood $p(x \mid \theta)$ directly from the data by semi-analytical calculation [17], such estimation has so far remained purely deterministic. Furthermore, the semi-analytical calculation of the likelihood is restricted to the exact model in (4). Including nonlinear parameters or changing the representation of the fluctuations to non-Gaussian or colored dynamics is not tractable with the semi-analytical method. Given the simplified structure of the physical model, a likelihood-free, amortized estimate of the posterior can have clear advantages over the likelihood-based estimate.

3 Experimental Set-up

In this section, we introduce the validation methods employed in our application and three baseline models for comparison with NPE in the following sections.

3.1 Training Details and Prior Choice

In order to train the neural posterior estimator/the MAF, we use 500k samples from the prior distributions, which are given in Table 1, and simulate a synthetic frequency trajectory with a length of 900 seconds and a time resolution of one second for each sample. A reliable choice for the interval length is crucial, as drifting parameters generating the dynamics for that interval can lead to uncertainty and a shift in the posterior estimates. Based on the division of the intraday market and the day-ahead market into 15-minute products, a sawtooth function is used to model the power imbalance. This represents the discrepancy between steady load and constant offered generation; therefore, the two parameters P_0 and P_1 can be assumed to be constant for quarter-hour intervals. Thus, this choice of interval provides a suitable scale for estimating the parameters. While a smaller interval size is suitable for analyzing short-term, noticeable frequency effects, it cannot sufficiently capture longer-term dynamics; conversely, the analysis of longer time windows is prone to overgeneralizing the estimated parameter values.

Table 1: Prior distribution of the starting values of the trajectory and the parameters.

Start value	Parameter	Prior distribution
$\phi(0)[\text{rad}]$		$\mathcal{U}(-1.25, 1.25)$
$\omega(0)[\text{rad/s}]$		$\mathcal{U}(-\pi, \pi)$
	$c_1[s^{-1}]$	$\mathcal{U}(0.001, 0.08)$
	$c_2[s^{-2}]$	$\mathcal{U}(0.00001, 0.0003)$
	$P_0[s^{-2}]$	$\mathcal{U}(-0.05, 0.05)$
	$P_1[s^{-3}]$	$\mathcal{U}(-0.00005, 0.00005)$
	$\epsilon[s^{-3/2}]$	$\mathcal{U}(0, 0.05)$

Note that we also sample the trajectory start values $\phi(0)$ and $\omega(0)$ from prior distributions to estimate the parameters of (4) with respect to trajectories of different start values. Therefore we refer to samples from the prior as $\theta = (\phi(0), \omega(0), c_1, c_2, P_0, P_1, \epsilon)$. When selecting the boundaries of prior distributions, we rely on point parameter estimates from the literature [17], and choose the boundaries such that the reference values fall within the prior distributions. To keep the constraints on the distributions as minimal as possible, we choose uniform distributions for the prior. It is relevant to choose a prior within an appropriate range of values, as a prior that is too narrow can truncate the true posterior. Conversely, a prior that is too broad can make the learning problem of the neural posterior estimator more elaborate, increasing training time and variance. Consequently, we define the lower and upper limits within a physically meaningful range to ensure that the trajectories generated with these parameters do not diverge.

3.2 Z-Scores and Calibration Tests

To evaluate our application of neural posterior estimation via masked autoregressive flows on synthetic data, we use three global and local validation methods that provide information about the quality of the inferred posterior. To apply these methods, we generate a set of prior samples, simulate a synthetic frequency trajectory for each sample, then infer the posterior and, again, draw a ensemble of samples from it, according to the following scheme:

(1) Sample $\theta \sim p(\theta)$.
(2) Simulate observed data $x_{\text{obs}} \sim p(x \mid \theta)$.
(3) Infer the posterior $q_\Theta(\theta \mid x_{\text{obs}})$.
(4) Draw posterior samples $\theta_{\text{post}} \sim q_\Theta(\theta \mid x_{\text{obs}})$.

Z-scores indicate the distance of a true parameter to the mean of the posterior samples, given in standard deviations from the mean:

$$z_i = \frac{\mu_{\text{post}} - \theta_i}{\sigma_{\text{post}}}, \tag{5}$$

where θ_i is the true parameter, while $\mu_{\text{post}} = \mathbb{E}(\theta_{\text{post}})$ and $\sigma_{\text{post}} = \sqrt{\text{Var}(\theta_{\text{post}})}$ are the mean and standard deviation of the inferred posterior distribution. We apply the calculation of z-scores to the five parameters of the aggregated swing equation and thus obtain an individual score for each parameter.

As another global diagnostic test, we apply simulation-based calibration (SBC), which provides a quantitative method to check whether the variances of the posterior are balanced [5, 10, 41]. SBC is a method for validating Bayesian inference algorithms by checking whether they can correctly recover parameters drawn from the prior when data are generated from the model itself. For each simulated dataset, the inference algorithm produces posterior samples, each associated with a log-probability representing its unnormalized posterior density. SBC defines the posterior rank of the true parameter value as the number of posterior samples that are smaller than the true parameter value. Intuitively, if inference is calibrated, the true parameter should be neither systematically more nor less probable than typical posterior samples, leading to uniformly distributed ranks. Aggregating many such ranks yields an empirical cumulative distribution function (ECDF), which should follow a straight diagonal if inference is correct. Deviations such as systematic slope changes or as curvature indicate biased, miscalibrated, or numerically unstable posterior computations.

3.3 Posterior Predictive Checks

Unlike with simulated data, we usually do not have access to the ground-truth posterior in empirical applications. Hence, in order to validate the estimated posterior distribution, we can instead rely on posterior predictive checks [10]. In the following, we give a systematic procedure for posterior predictive checks:

(1) Sample $\theta \sim p(\theta)$.
(2) Simulate observed data $x_{\text{obs}} \sim p(x \mid \theta)$.
(3) Infer the posterior $q_\Theta(\theta \mid x_{\text{obs}})$.
(4) Draw posterior samples $\theta_{\text{post}} \sim q_\Theta(\theta \mid x_{\text{obs}})$.
(5) For each posterior draw, generate posterior predictive data $x_{\text{post}} \sim p(x \mid \theta_{\text{post}})$.
(6) Compare x_{obs} with the replicated data $\{x_{\text{post}}\}$ both qualitatively and quantitatively.

For applying NPE on empirical data, steps (1) and (2) are skipped, as we infer the posterior directly from the real observation x_{obs}. In order to quantitatively evaluate the similarity of the simulations from the inferred posterior with the observed data, we need model scores. Here, we address the overall accuracy of the ensemble of simulations, as well as the accuracy of their mean.

Energy Score The intention of sampling from the posterior and further creating simulations is to compare these synthetic observations with empirical observations. If we apply different parameter samples from the posterior distribution in the simulator, we obtain an ensemble of trajectories, in contrast to a point estimate. In order to properly evaluate the quality of the simulations with respect to the true observations, we make use of the energy score

$$ES(X, y) = \frac{1}{N} \sum_{1}^{N} \left(\left\| X^{(i)} - y \right\|_2 \right) - \frac{1}{2N^2} \sum_{i,j}^{N} \left(\left\| X^{(i)} - X^{(j)} \right\|_2 \right) \quad (6)$$

for an ensemble X of segments $X^{(i)}$ and observation y.

Mean absolute error between empirical observation and mean of simulations from posterior. In addition to measuring the fit of the entire ensemble of simulations, we also compare the mean of the simulations to the empirical observation by calculating the mean absolute error (MAE) to the observation.

3.4 Baseline Models

To classify validation results of posterior predictive samples, simulated by the inferred posterior, we compare the simulations generated by the inferred posterior to three baseline methods. For this purpose, we choose one parameter-free method for the empirical case that compares historical time series segments with the observed data, a basic method of simulation-based inference based purely on prior sampling, data simulation, and comparison between simulated and observed data, as well as a deterministic estimation using least square estimates.

3.4.1 *Daily pattern.*

Frequency dynamics are dependent on the grid's generation mix and the current power demand. Available generation types as well as the load are influenced by calendar features, such as the time of day, the day of the week, and the current season. In order to create a simple, purely frequency-dependent baseline that is capable of simulating frequency trajectories and comparing them to empirical observations, we use the ensemble of frequency trajectories from the year before at the same time of the day as a baseline model: If we consider a specific quarter of an hour as an observation segment, we sample 100 unweighted 15-minute trajectories from frequency measurements from the year before at the same time of the day. An algorithmic description of the method can be found in the Appendix A. In this way, we obtain an approximation for the relevant frequency segment from a set of measurements taken at the same time of day in the past. Note that, as we are using data from 2025 for parameter estimation, we are using baseline data from 2024 for sampling from the daily pattern.

3.4.2 *ABC rejection [34].*

Density estimation methods such as NPE, which we apply in the present paper, and further neural density estimation and neural ratio estimation, are based on deep neural networks. Approximate Bayesian Computation (ABC) [36] represents a simplified method of simulation-based inference: We draw samples from the prior distribution $p(\theta)$, and simulate the model with these parameters. In the following, we use the Euclidean distance to measure the mismatch between the summary statistics of the simulation and the observation. The summary statistics contain fourteen features including standardized moments of the trajectories; the full description of the summary statistics can be found in Appendix C. If the value of the metric lies within a selected percentile of all simulations, the parameters of the simulation get accepted and therefore are added to the posterior distribution; otherwise, the prior sample gets rejected. In our case, to be consistent with the NPE method, we simulate trajectories for 500k prior samples and keep the share of prior samples for which the resulting trajectory summary statistics are below the 0.1% percentile of all samples as our ABC posterior.

3.4.3 *Deterministic least-square estimate (LSE) with SINDy.*

Having established two probabilistic baseline models, we include a third, deterministic parameter estimation baseline based on the least square estimate [21] of the model parameters. We apply a Gaussian kernel filter with a 10-second window to the time series segments under consideration to remove noise from the signal and receive ω_{trend}. For this denoised signal, we then use least-squares estimation with the implementation of Sparse Identification of Nonlinear Dynamics(SINDy) [4, 7], a model for sparse estimation of regression parameters, to estimate the parameters for the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\omega = \hat{f}_{\text{LSE}}(\omega, \phi, t) = \Theta(\phi, \omega, t)\Xi,$$

where the function library is given by

$$\Theta(\phi, \omega, t) = (\omega, \phi, 1, t),$$

and the parameters

$$\Xi = (c_1, c_2, P_0, P_1)^T$$

are estimated via least squares. In order to estimate the noise amplitude, we compute the empirical variance of the residual increments as

$$\epsilon = \left(\frac{1}{n\Delta t}\sum_k \left[(\omega_{k+1} - \omega_k) - \Delta t \hat{f}_{\text{LSE}}(\omega_k, \phi_k, t_k)\right]^2\right)^{\frac{1}{2}},$$

with $\Delta t = 1$ s and n the number of samples in a 15-minute window. Note that, as the LSE baseline is defined as a point estimate, we run the model with different noise realizations in order to receive an ensemble of trajectories for the posterior predictive checks, i.e. to calculate the energy score and the MAE. This differs from ABC and NPE, where we sample from the parameter distribution to generate trajectories.

4 Validation on Synthetic Dataset

Before we apply the NPE approach to empirical frequency data, we validate the performance of NPE in our proposed application on a purely simulated data set, where ground-truth parameters are available. For the validation methods introduced in 3.2 and 3.3, we draw, with the exception of the calibration test, 1000 parameter samples from the prior distribution and simulate frequency trajectories. After inferring the posterior, we draw 100 samples from each posterior distribution. This allows us to perform diagnostics for the inferred posterior and evaluate simulations with parameters from the posterior distribution. The z-scores for the NPE posterior distributions, as shown in Fig. 2a, are approximately Gaussian distributed, which indicates that the posterior has correct coverage properties and the uncertainty is neither under- nor overestimated.

By applying simulation-based calibration for NPE, we calculate the empirical cumulative distribution function of the posterior ranks, shown in Fig. 2b. Note that here, unlike with the other validation methods, we draw 200 prior samples and 1000 samples each from the posterior distributions to achieve a sufficiently accurate approximation of the uniform rank distribution and an adequate resolution of the rank statistics. The observed rank distribution approximately follows the diagonal, indicating that the posterior appears well calibrated and does not exhibit systematic miscalibration.

We now turn to the analysis of the simulations with parameters from the posterior, using the method of posterior predictive checks, and compare the performance of NPE, ABC, and LSE. To evaluate

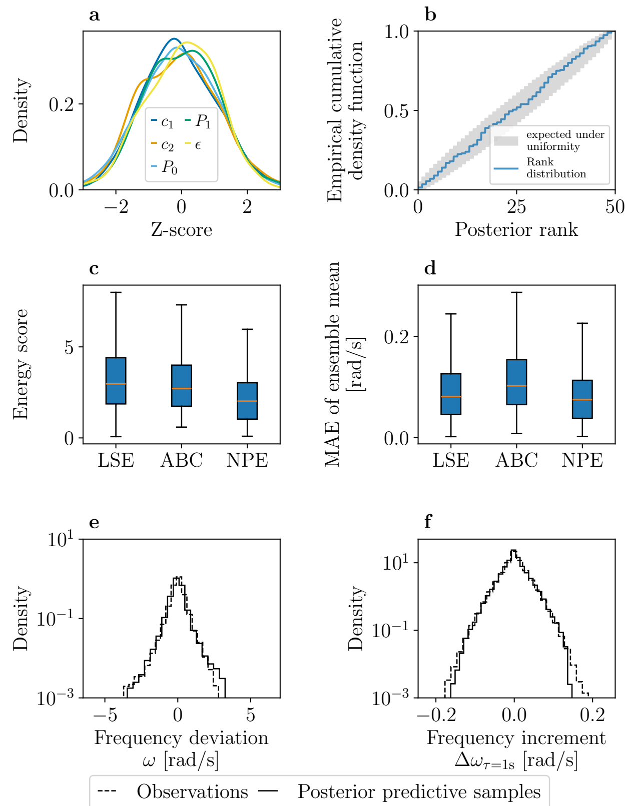

Figure 2: Validation of the neural posterior estimator on simulated observations. a: Z-scores of posterior samples relative to the ground-truth parameter. b: Simulation-based calibration using comparison of posterior ranks and empirical cumulative distribution function. c and d: Comparison of LSE, ABC, and NPE posterior predictive samples with respect to the energy score and the MAE of the posterior predictive ensemble mean. e: Probability density functions of the synthetic observations and posterior predictive samples. f: Probability density functions for the time series increments of the synthetic observations and posterior predictive samples.

the general performance of NPE for the aggregated swing equation, we use the energy score as well as the MAE of the ensemble mean, and compare the results of NPE with those of the ABC and LSE baselines. In Fig. 2c and d, we note that NPE achieves better results with respect to both scores. With regard to the narrow margin of non-rejected parameters in the ABC baseline, this observation shows us that, in terms of posterior predictive checks, NPE shows promising results. We observe that LSE performs nearly as well as NPE in terms of the MAE, while it performs worse in terms of the energy score. This suggests an overestimation of the noise amplitude in LSE.

In the following, we compare the posterior predictive samples with the synthetic observations by analyzing the probability distributions of the samples. To achieve this, we generate a posterior predictive sample for each of the 1000 samples and compare the probability density functions (PDF) of the entire synthetic observations and posterior predictive samples. In addition, we also calculate the PDF of the frequency increments, which are given as

$$\Delta\omega_\tau = \omega(t+\tau) - \omega(t),$$

and describe the differences between data points in a time series at time intervals of length τ. Fig. 2e and f show the PDF of the frequency time series segments respective its increments with time lag $\tau = 1$ s. We observe very similar behavior of synthetic observations and posterior predictive samples with regard to both density estimates.

Due to the influence of highly volatile renewable generation, power systems are often affected by the occurrence of temporally correlated fluctuations. To quantify the robustness of our approach, we therefore evaluate the performance of parameter inference on frequency time series segments with fractional Gaussian noise [25] of varying degrees of correlation. This demonstrates how the posterior estimator, trained for time series with uncorrelated fluctuations, performs for time series segments with correlated fluctuations. In the analysis presented in Appendix D and Figure 8, we observe that only correlated noise with a Hurst exponent larger than 0.6 has a considerable impact on the calibration quality of the posterior parameters, with the exception of the noise amplitude. We conclude that fluctuations with negative or weak positive correlations do not have a substantial negative impact on the quality of inference.

5 Parameter Estimation on Empirical Frequency Time Series from European Synchronous Areas

After successfully validating neural posterior estimation for parameter inference of the aggregated swing equation on simulations generated from prior samples, we now apply the trained model to empirical frequency data. As before, we consider time series segments of 15 minutes. For model comparison with the three baselines, we observe 1000 samples of frequency measurements from 2025 from the Continental European and the Nordic synchronous areas, the latter comprising Sweden, Norway, Finland and the eastern part of Denmark. For each sample x_{obs}, we draw 100 parameter samples from the posterior $\theta \sim q_\Theta(\theta \mid x_{\text{obs}})$, and generate for each of them a posterior predictive sample, i.e. one synthetic frequency simulation. Hence, for each empirical time series segment, we obtain an ensemble of posterior predictive samples.

5.1 Model Performance

To assess how well the different models (NPE and the three baseline models) capture the empirical frequency data, we compute the energy score and the MAE between the mean of the ensemble of posterior predictives and the observation. For both the CE and the Nordic grid, NPE posterior clearly outperforms the baseline models in both metrics, which is shown in Fig. 3. This means that the parameters sampled from the posterior distribution of the NPE model lead to trajectories that approximate the observation better than sampling from the daily pattern or using ABC or LSE. This is remarkable given that the daily profile is typically a very good approximation already [17].

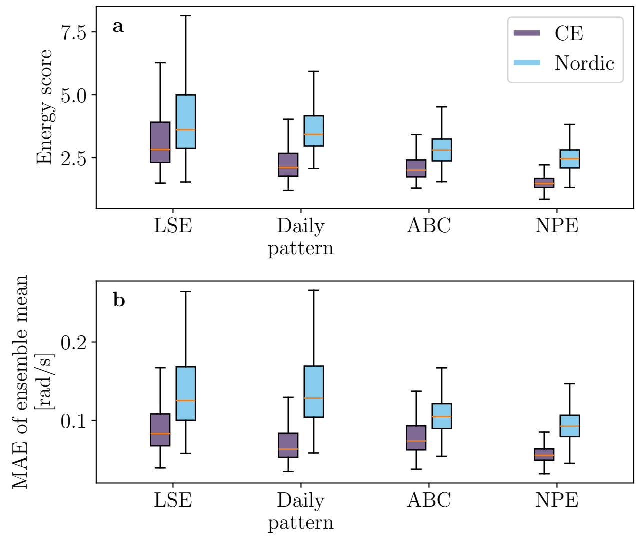

Figure 3: Distribution of performance metrics across 1000 empirical frequency time series segments from 2025 in the Continental European (CE) and the Nordic synchronous area. Energy score (a) and mean absolute error (MAE) of the ensemble mean (b) are shown for four models (LSE, Daily Pattern, ABC, and NPE) as boxplots.

5.2 Analysis of Posterior Predictive Trajectories

Having established that NPE is performing well on the whole data set, we analyze the inferred posterior by inspecting individual trajectories. This allows a qualitative judgment of how well the inferred dynamics capture key dynamic properties of the empirical data set.

Specifically, we evaluate posterior simulations using the energy score and analyze empirical samples together with the corresponding posterior simulations that achieve the best and worst scores across the full CE test set of 1000 samples, see Fig. 4.

For the segment with the best energy score under NPE, see Fig. 4a-d, the trajectories simulated with samples from the inferred NPE posterior fit the empirical time series very well. The shaded area, representing the standard deviation of the simulations, covers the trajectory of the empirical sample for most of the period. Further, the colored line, which indicates the mean value of the simulation ensemble, is centered in the vertical extension of the trajectory. It is noteworthy that within the first minute of the frequency trajectory, the variance of the simulation ensemble is very low. If we consider the same trajectory for the baseline models, we see that simulations generated by the estimated ABC posterior do not reproduce the positive upward trend of the trajectory after the first few minutes. The parameter-free daily pattern model, which approximates the trajectory using sampled trajectories from the previous year at the same time of day, again matches the trend of the trajectory, but less accurately and with significantly greater variance. Further, the LSE baseline, which is based on point estimates and constructed using various realizations of Gaussian noise, also follows the frequency response on average. However, it has a significantly higher noise

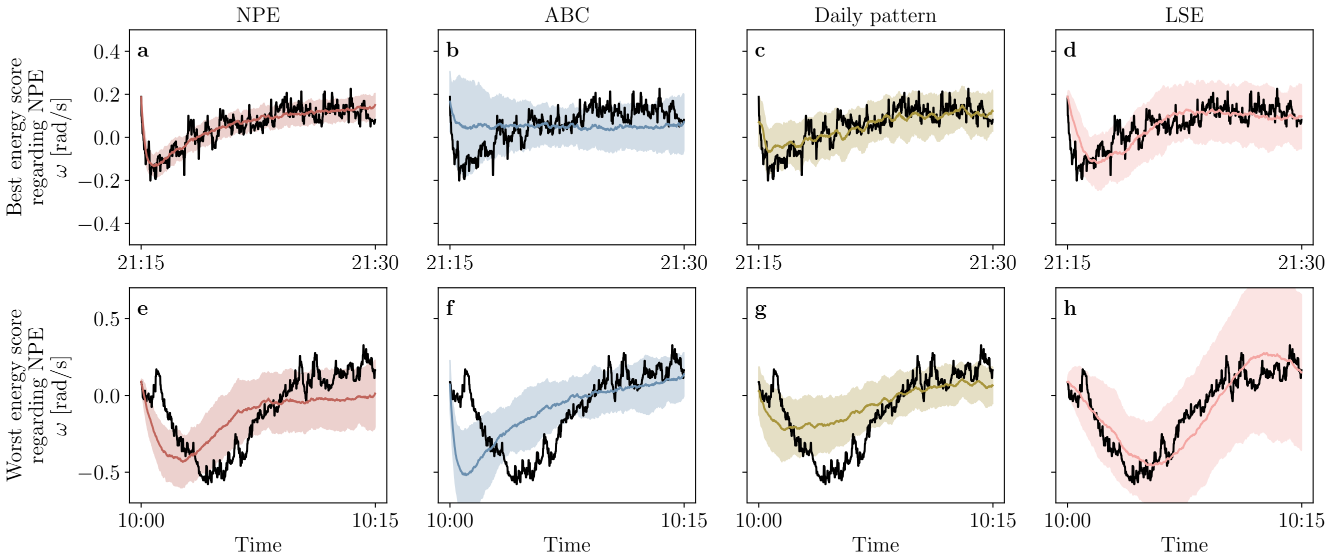

Figure 4: Comparison of observed CE frequency segment trajectories simulated by 100 samples from the posterior. Black curves are the empirical frequency segments, shaded areas indicate the standard deviation of the trajectories, and colored lines represent the mean of the ensemble of trajectories. a-d: Segment with the best energy score with respect to the NPE model, with the corresponding trajectories of all four models. e-h: Segment with the worst energy score with respect to the NPE model, with the corresponding trajectories of all four presented models.

level than NPE and does not fully capture the drop at the start of the hour.

If we now look at the period with the worst energy score in terms of NPE in Fig. 4e-h, we observe that none of the three models - NPE, ABC, and Daily Pattern - can accurately predict the true trajectory over the entire 15-minute period. We identify the reason why NPE and ABC cannot approximate the trajectory, particularly at the beginning, due to frequency deviation shortly after the start of the period. This deviation, which does not occur at the full hour, is not specified by the construction of the aggregated swing equation and therefore cannot be simulated using it. While all these three models fit the trend of the frequency trajectory after the first few minutes, we notice that the simulation ensemble by NPE is more widely distributed than the samples of ABC and Daily Pattern, i.e. it at least reports a high uncertainty when it performs poorly. The problem of model misspecification with regard to the time series segment under consideration poses a challenge to extend the simulator to such cases and, accordingly, to process a larger selection of parameters. In contrast, the ensemble of trajectories generated by LSE covers the empirical frequency trajectory over the entire interval; however, the variance increases noticeably over the time interval.

Analysis of the best and worst predictions of NPE for Nordic frequency data shows qualitatively similar results; for details, see Appendix E and Fig. 9.

In addition, as for the validation of NPE on synthetic frequency data segments, we calculate the PDFs of empirical data as well as of posterior predictive samples, generated from the inferred posterior distribution. As we can see from the distribution of frequency trajectories in Fig. 5a, the distribution of empirical data (solid lines) and predictions (dashed lines) coincide strongly in both synchronous areas. Regarding the 1-second increments in Fig. 5b, the forecasts slightly underestimate the variance in Continental Europe, while a more significant overestimation is noticeable in the Nordic system. This suggests that the Nordic grid is subject to complex fluctuation dynamics that cannot be accurately approximated with uncorrelated Gaussian noise. However, if we consider 10-second increments, as seen in Fig. 5c, we see that the distributions of observations and predictions in both networks exhibit very similar variance. This is an indicator that parameter inference with NPE is capable of modeling dynamics on a higher time scale even in the Nordic grid, which has more complex stochastic dynamics.

5.3 Temporal Trends in Posterior Distributions

While we assume the prior distributions to be independent of time and each other, we observe significant influences of the respective time of day on the frequency dynamics in the daily profile of the frequency time series. By inferring only short segments of fifteen minutes of frequency data, we enable the observation of time- and frequency-specific effects on posterior distributions.

In Fig. 6, we show the daily profile of the means and standard deviations of the inferred posterior distributions for CE. Daily profile refers to the average of every quarter hour of a day throughout the entire year. Significant fluctuations at the beginning of each hour are noticeable in all parameter distributions, both regarding the mean and the standard deviation. This is caused by significant frequency deviations at the beginning of each hour, which in turn

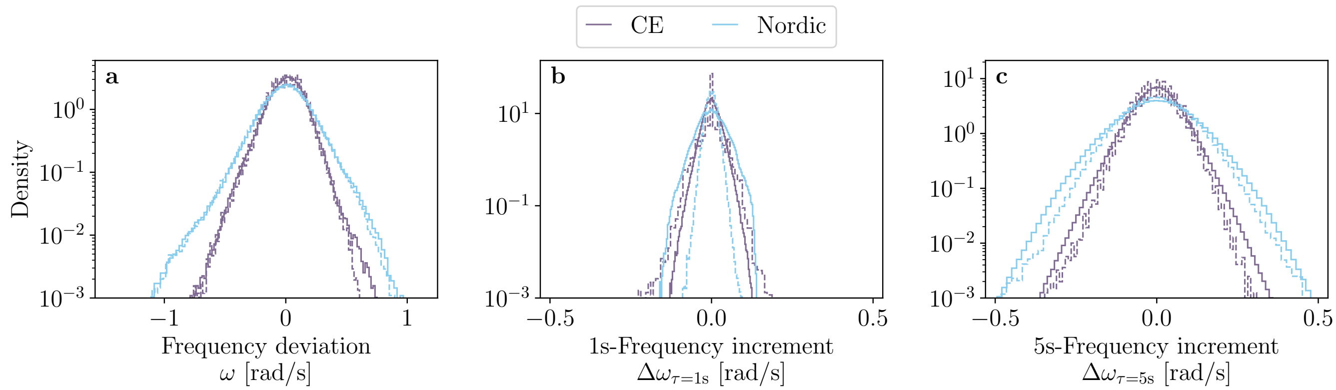

Figure 5: Comparison of the frequency and frequency increments distributions for empirical observations and simulated samples from the posterior. Dashed lines indicate the empirical time series, solid lines represent trajectories that are simulated with samples from the inferred posterior.

are caused by significant changes in generation dispatch at the start of each hour [46].

Beyond that, each of the five parameter distributions shows a clear daily pattern when observing its mean, as shown in Fig. 6a-e. Primary (c_1) and secondary (c_2) control follow a similar trend, with peaks between 00:00 and 06:00 and between 12:00 and 15:00, and lows between 06:00 and 12:00 and between 15:00 and 18:00. The relation $P_0 + P_1\ t$ of the parameters P_0 and P_1, which represents the imbalance between scheduled generation and continuous load, forms a sawtooth curve with a negative intercept and positive slope around 06:00 and 18:00, which qualitatively corresponds to the behavior of generation and load. This observation is consistent with results reported in the literature [17]. Further, P_0 and P_1, see Fig. 6c and d, indicate weaker fluctuations of the mean than the other parameters. This can be explained by the regularity of deterministic frequency fluctuations. The mean values of the noise amplitude distribution in Fig. 6e show a clear difference between day and night, with a consistently high level during the day.

In addition to the trend of the distribution means, we further analyze the width of the distribution around the mean, i.e. the standard deviation, during the course of a day in Fig. 6f-j. While the mean of the distributions for all parameters, and thus the distribution itself, shows clear shifts, the width of the inferred posterior distributions also shows clear daily trends. It is evident that each parameter shows a correlation between the size of the standard deviation and the mean value. We will turn to a more detailed analysis of the behavior of the variance of the parameter distributions in 5.4.

Analysis of the Nordic grid shows similar qualitative behavior of the parameter means and standard deviations in the daily profile, however, with less significant deviations throughout the day, see Section F.

5.4 Correlation between Observations and Posterior Distribution Variance

In contrast to point estimation by using neural posterior estimation, we receive information about the certainty of the estimated parameter values, derived from the properties of the posterior distributions. After having observed fluctuations in the distribution shape throughout the daily profile, we analyze the relation between the properties of the observed time series from the CE synchronous area and the behavior of the inferred posterior parameters. In order to quantitatively describe that relation, we consider the standard deviation of the time series segment and the standard deviation of the estimated parameter distribution for 1000 empirical samples, the latter approximated from 100 posterior samples. We observe significant negative correlations between the variance of the observation and the standard deviation of the control parameters c_1 and c_2 as shown in Fig. 7a and b. This indicates that, for time series segments that exhibit large fluctuations from the reference frequency, the posterior estimation for c_1 and c_2 is significantly narrower than for samples with smaller deviations. We notice a different effect for the imbalance parameter P_0 and P_1 in Fig. 7, where the correlation between the variances of the observation and the parameters is positive. However, with respect to noise amplitude ϵ in 7e, we do not observe any correlation. The observations lead us to conclude that the properties of the observed time series segments affect the variance of the distributions, and thus the selection of parameters that are considered realistic for the observation. For comparability, we analyze the frequency data from the Nordic grid using the same method and make similar qualitative observations, see Fig. 11 in the Appendix.

6 Conclusion and Future Work

In the present paper, we apply simulation-based inference using neural posterior estimation for power grid frequency time series. Validating the model for the proposed prior distributions and the underlying physical model shows that this amortized inference estimation technique is suitable for our use case. In addition to validation on synthetic data based on parameters sampled from the prior, we apply the model to empirical frequency data from two European synchronous areas and infer parameter distributions for the underlying physical model. Due to the absence of ground-truth parameters in the empirical case study, we demonstrate the model

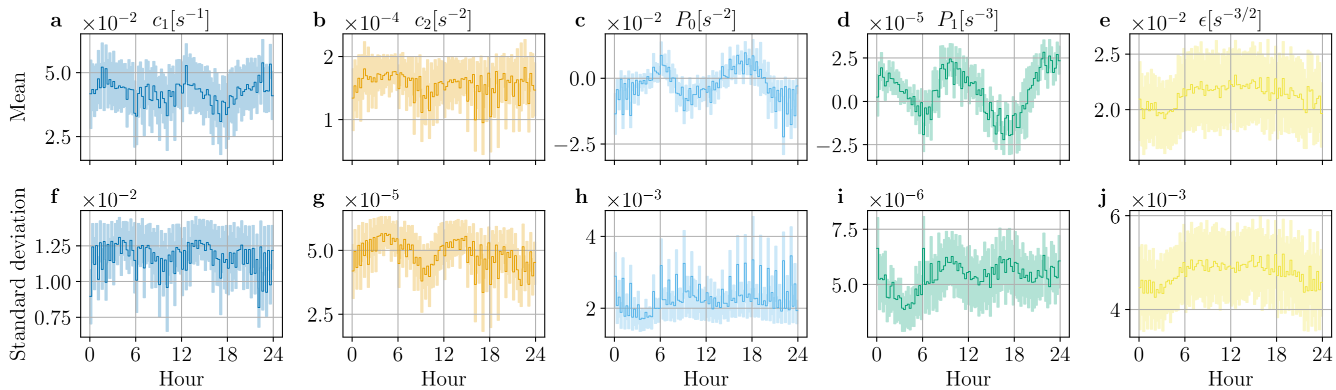

Figure 6: Daily patterns in mean (a-e) and standard deviation (f-j) of the posterior distribution of the model parameters in CE. Solid lines represent the daily means, and shaded areas indicate the range between the 25% and 75% quantiles.

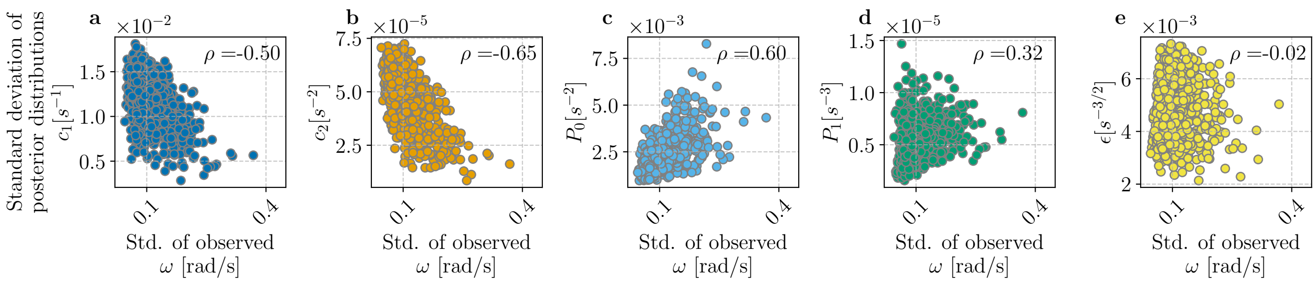

Figure 7: Relation between the standard deviation of CE frequency time series segments and the deviation of the corresponding parameters from the inferred posterior. The Pearson correlation coefficient between the corresponding standard deviations is given by ρ.

performance based on the evaluation of posterior predictive checks and outperform three baseline models. Further, we analyze the inferred posterior parameter distributions and investigate trends and correlations in the distributions.

Empirical stochastic time series have been found to be an interesting use case for simulation-based inference. This posterior estimate can now be applied to the Continental European grid or other systems to infer control or noise parameters, enabling more informed balancing actions, such as reducing control effort to lower costs or increasing it to enhance frequency quality.

By estimating the posterior distribution segment-wise for 15-minute intervals, we demonstrate the significant time-dependent dynamics of the estimated parameters. Furthermore, we observe correlations between empirical observations and the corresponding posterior distributions.

Thus, neural posterior estimation proves to be a relevant analysis technique for time series data and, in particular, for power grid frequency. However, neural posterior estimation can also play a relevant role in other areas of application for energy systems.

Having established neural posterior density estimation for the aggregated swing equation (4), this opens numerous future research directions. For example, the noise term could be modeled more realistically. Specifically, both correlated, i.e. non-white, but also non-Gaussian noise could be considered. This would further make NPE useful as power systems might be strongly correlated or display non-Gaussian fluctuations. As we show for frequency time series from the Nordic grid, implementing more complex noise dynamics in the simulator may increase the performance of NPE. For this, the application of a likelihood-free posterior inference method enables an expansion of the model towards nonlinearity and non-Gaussianity without restrictions. Further, adaptations in the model architecture of the neural density estimation may be able to address the discrepancy between the physical model and empirical power grid frequency time data. Recently, robust methods have been applied to adapt simulation-based inference under model misspecification [14, 29]. Hence, the combination of more complex neural density estimators and a more sophisticated simulator architecture can offer useful advances.

Estimating central power systems variables and their distribution is important in designing and adapting power-grid control mechanisms to ensure a robust power supply. Primary control strength can reveal various weaknesses of the ability of a power system to quickly and adequately return to its nominal state. The power balance needs constant monitoring, and the ability to estimate their parametrization permits system operators to prepare preemptive control. As more renewable energies enter our power system's composition, noise levels are likely to increase to potentially dangerous values, and primary inertial control will weaken. Assessing these parameters efficiently can thus serve both as good diagnostic tools as well as fast, reliable control variables to be used in the system's operation.

Acknowledgments

We gratefully acknowledge funding from the Helmholtz Association and the Networking Fund through Helmholtz AI and under grant no. VH-NG-1727. We also acknowledge support by the state of Baden-Württemberg through bwHPC. This work was further funded from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 556503410.

Data and Code Availability

Frequency data with a resolution of one second on transmission grid level for the CE and Nordic synchronous areas is provided by the transmission system operators [9, 26]. The code for reproducing the parameter inference is available at [28].

References

[1] Mark A Beaumont, Wenyang Zhang, and David J Balding. 2002. Approximate Bayesian Computation in Population Genetics. *Genetics* 162, 4 (Dec. 2002), 2025–2035. doi:10.1093/genetics/162.4.2025

[2] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational Inference: A Review for Statisticians. *J. Amer. Statist. Assoc.* 112, 518 (April 2017), 859–877. doi:10.1080/01621459.2017.1285773

[3] Jan Boelts, Michael Deistler, Manuel Gloeckler, Álvaro Tejero-Cantero, Jan-Matthis Lueckmann, Guy Moss, Peter Steinbach, Thomas Moreau, Fabio Muratore, Julia Linhart, Conor Durkan, Julius Vetter, Benjamin Kurt Miller, Maternus Herold, Abolfazl Ziaeemehr, Matthijs Pals, Theo Gruner, Sebastian Bischoff, Nastya Krouglova, Richard Gao, Janne K. Lappalainen, Bálint Mucsányi, Felix Pei, Auguste Schulz, Zinovia Stefanidi, Pedro Rodrigues, Cornelius Schröder, Faried Abu Zaid, Jonas Beck, Jaivardhan Kapoor, David S. Greenberg, Pedro J. Gonçalves, and Jakob H. Macke. 2025. Sbi Reloaded: A Toolkit for Simulation-Based Inference Workflows. arXiv:2411.17337 [cs] doi:10.48550/arXiv.2411.17337

[4] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. 2016. Discovering Governing Equations from Data by Sparse Identification of Nonlinear Dynamical Systems. *Proceedings of the National Academy of Sciences* 113, 15 (April 2016), 3932–3937. doi:10.1073/pnas.1517384113

[5] Samantha R Cook, Andrew Gelman, and Donald B Rubin. 2006. Validation of Software for Bayesian Models Using Posterior Quantiles. *Journal of Computational and Graphical Statistics* 15, 3 (Sept. 2006), 675–692. doi:10.1198/106186006X136976

[6] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. 2020. The Frontier of Simulation-Based Inference. *Proceedings of the National Academy of Sciences* 117, 48 (Dec. 2020), 30055–30062. doi:10.1073/pnas.1912789117

[7] Brian M. de Silva, Kathleen Champion, Markus Quade, Jean-Christophe Loiseau, J. Nathan Kutz, and Steven L. Brunton. 2020. PySINDy: A Python Package for the Sparse Identification of Nonlinear Dynamical Systems from Data. *Journal of Open Source Software* 5, 49 (May 2020), 2104. doi:10.21105/joss.02104

[8] Conor Durkan, Iain Murray, and George Papamakarios. 2020. On Contrastive Learning for Likelihood-free Inference. *Proceedings of the 37th International Conference on Machine Learning (ICML), Vol. 37* (2020), 2771–2781.

[9] Fingrid. 2026. Frequency data for the *nordic* grid. https://data.fingrid.fi/en/datasets/339. Accessed: 2026-01-10.

[10] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. Visualization in Bayesian Workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society* 182, 2 (Feb. 2019), 389–402. doi:10.1111/rssa.12378

[11] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. 2015. MADE: Masked Autoencoder for Distribution Estimation. In *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, Lille, France, 881–889.

[12] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. 2019. Automatic Posterior Transformation for Likelihood-Free Inference. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2404–2414. https://proceedings.mlr.press/v97/greenberg19a.html

[13] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. 2022. A Crisis In Simulation-Based Inference? Beware, Your Posterior Approximations Can Be Unfaithful. *Transactions on Machine Learning Research* (Sept. 2022).

[14] Daolang Huang, Ayush Bharti, Amauri H. Souza, Luigi Acerbi, and Samuel Kaski. 2023. Learning Robust Statistics for Simulation-Based Inference under Model Misspecification. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, 7289–7310.

[15] Tommi S. Jaakkola and Michael I. Jordan. 2000. Bayesian Parameter Estimation via Variational Methods. *Statistics and Computing* 10, 1 (Jan. 2000), 25–37. doi:10.1023/A:1008932416310

[16] Patrick Kidger. 2021. *On Neural Differential Equations*. Ph. D. Dissertation. University of Oxford.

[17] Johannes Kruse, Eike Cramer, Benjamin Schäfer, and Dirk Witthaut. 2023. Physics-Informed Machine Learning for Power Grid Frequency Modeling. *PRX Energy* 2, 4 (Oct. 2023), 043003. doi:10.1103/PRXEnergy.2.043003

[18] Johannes Kruse, Benjamin Schäfer, and Dirk Witthaut. 2021. Exploring deterministic frequency deviations with explainable AI. In *2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. 133–139. doi:10.1109/SmartGridComm51999.2021.9632335

[19] Amanda Lenzi, Julie Bessac, and Mihai Anitescu. 2022. Detecting Large Frequency Excursions in the Power Grid With Bayesian Decision Theory. *IEEE Open Access Journal of Power and Energy* 9 (2022), 66–75. doi:10.1109/OAJPE.2022.3144612

[20] Lennart Ljung. 1999. *System Identification: Theory for the User* (2 ed.). Prentice Hall, Upper Saddle River, NJ.

[21] Lennart Ljung. 2010. Perspectives on system identification. *Annual Reviews in Control* 34, 1 (2010), 1–12.

[22] Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H. Macke. 2019. Likelihood-Free Inference with Emulator Networks. arXiv:1805.09294 [stat] doi:10.48550/arXiv.1805.09294

[23] Jan-Matthis Lueckmann, Pedro J. Gonçalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H. Macke. 2017. Flexible statistical inference for mechanistic models of neural dynamics. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 1289–1299.

[24] Hannan Ma and Husheng Li. 2013. Analysis of Frequency Dynamics in Power Grid: A Bayesian Structure Learning Approach. *IEEE Transactions on Smart Grid* 4, 1 (March 2013), 457–466. doi:10.1109/TSG.2012.2226066

[25] Benoit B. Mandelbrot and John W. Van Ness. 1968. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Rev.* 10, 4 (1968), 422–437. http://www.jstor.org/stable/2027184

[26] Netztransparenz. 2026. Frequency data for the *CE* grid. https://www.netztransparenz.de/de-de/Regelenergie/Daten-Regelreserve/. Accessed: 2026-01-10.

[27] Alexandra Nikoltchovska, Sebastian Pütz, Xiao Li, Veit Hagenmeyer, and Benjamin Schäfer. 2025. Probabilistic and Explainable Machine Learning for Tabular Power Grid Data. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '25)*. Association for Computing Machinery, New York, NY, USA, 213–231. doi:10.1145/3679240.3734623

[28] Ulrich Oberhofer, Nicolas Weber, Veit Hagenmeyer, and Benjamin Schäfer. 2026. NPE-for-grid-frequency. https://github.com/KIT-IAI-DRACOS/NPE-for-grid-frequency.

[29] Matthew O'Callaghan, Kaisey S. Mandel, and Gerry Gilmore. 2025. Misspecification-Robust Amortised Simulation-Based Inference Using Variational Methods. arXiv:2509.05724 [stat] doi:10.48550/arXiv.2509.05724

[30] George Papamakarios and Iain Murray. 2016. Fast -Free Inference of Simulation Models with Bayesian Conditional Density Estimation. (2016), 1028–1036.

[31] George Papamakarios, Theo Pavlakou, and Iain Murray. 2017. Masked Autoregressive Flow for Density Estimation. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 2338–2347.

[32] George Papamakarios, David Sterratt, and Iain Murray. 2019. Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. PMLR, 837–848.

[33] Danilo Rezende and Shakir Mohamed. 2015. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, 1530–1538.

[34] Donald B. Rubin. 1984. Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics* 12, 4 (Dec. 1984), 1151–1172. doi:10.1214/aos/1176346785

[35] Leonardo Rydin Gorjão, Richard Jumar, Heiko Maass, Veit Hagenmeyer, G. Cigdem Yalcin, Johannes Kruse, Marc Timme, Christian Beck, Dirk Witthaut, and Benjamin Schäfer. 2020. Open Database Analysis of Scaling and Spatio-Temporal Properties of Power Grid Frequencies. *Nature Communications* 11, 1 (Dec. 2020), 6362. doi:10.1038/s41467-020-19732-7

[36] S. A. Sisson, Y. Fan, and M. A. Beaumont. 2018. Overview of Approximate Bayesian Computation. arXiv:1802.09720 [stat] doi:10.48550/arXiv.1802.09720

[37] Oliver Smith, Oliver Cattell, Etienne Farcot, Reuben D. O'Dea, and Keith I. Hopcraft. 2022. The Effect of Renewable Energy Incorporation on Power Grid Stability and Resilience. *Science Advances* 8, 9 (March 2022), eabj6734. doi:10.1126/sciadv.abj6734

[38] Simon Stock, Davood Babazadeh, Christian Becker, and Spyros Chatzivasileiadis. 2024. Bayesian Physics-informed Neural Networks for System Identification of Inverter-Dominated Power Systems. *Electric Power Systems Research* 235 (Oct. 2024), 110860. doi:10.1016/j.epsr.2024.110860

[39] Simon Stock, Jochen Stiasny, Davood Babazadeh, Christian Becker, and Spyros Chatzivasileiadis. 2023. Bayesian Physics-Informed Neural Networks for Robust System Identification of Power Systems. In *2023 IEEE Belgrade PowerTech*. 1–6. doi:10.1109/PowerTech55446.2023.10202692

[40] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. 2013. Approximate Bayesian Computation. *PLOS Computational Biology* 9, 1 (Jan. 2013), e1002803. doi:10.1371/journal.pcbi.1002803

[41] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. 2020. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv:1804.06788 [stat] doi:10.48550/arXiv.1804.06788

[42] Andreas Ulbig, Theodor S. Borsche, and Göran Andersson. 2014. Impact of Low Rotational Inertia on Power System Stability and Operation. *IFAC Proceedings Volumes* 47, 3 (2014), 7290–7297. doi:10.3182/20140824-6-ZA-1003.02615

[43] Petr Vorobev, David M Greenwood, John H Bell, Janusz W Bialek, Philip C Taylor, and Konstantin Turitsyn. 2019. Deadbands, droop, and inertia impact on power system frequency distribution. *IEEE Transactions on Power Systems* 34, 4 (2019), 3098–3108. doi:10.1109/TPWRS.2019.2895547

[44] Lei Wang and Junjian Qi. 2025. Approximate Bayesian Computation Guided Sequential Neural Posterior Estimation for Power System Parameter Calibration. *IEEE Transactions on Power Systems* (2025), 1–12. doi:10.1109/TPWRS.2025.3647318

[45] Daniel Ward, Patrick Cannon, Mark Beaumont, Matteo Fasiolo, and Sebastian M. Schmon. 2022. Robust Neural Posterior Estimation and Statistical Model Criticism. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, 33845–33859.

[46] T. Weissbach and E. Welfonder. 2009. High Frequency Deviations within the European Power System: Origins and Proposals for Improvement. In *2009 IEEE/PES Power Systems Conference and Exposition*. 1–6. doi:10.1109/PSCE.2009.4840180

[47] Allen J Wood, Bruce F Wollenberg, and Gerald B Sheblé. 2013. *Power generation, operation, and control.* John wiley & sons.

[48] P. Zarco and A.G. Exposito. 2000. Power System Parameter Estimation: A Survey. *IEEE Transactions on Power Systems* 15, 1 (Feb. 2000), 216–222. doi:10.1109/59.852124

[49] Junbo Zhao, Antonio Gómez-Expósito, Marcos Netto, Lamine Mili, Ali Abur, Vladimir Terzija, Innocent Kamwa, Bikash Pal, Abhinav Kumar Singh, Junjian Qi, Zhenyu Huang, and A. P. Sakis Meliopoulos. 2019. Power System Dynamic State Estimation: Motivations, Definitions, Methodologies, and Future Work. *IEEE Transactions on Power Systems* 34, 4 (July 2019), 3188–3198. doi:10.1109/TPWRS.2019.2894769

A Description of the Daily Pattern

We construct the Daily Pattern baseline as follows:

Algorithm 1: Daily Pattern Ensemble Construction

Data: 15-minute frequency time series segment; historical data from the previous year
Result: Ensemble of frequency time series segments
Determine corresponding quarter-hour index of the time series segment;
Extract all segments from the previous year with the same quarter-hour index;
Uniformly sample N segments without weighting;
Construct the ensemble from the sampled segments;
return Ensemble of frequency segments

B Computational Effort and Scalability

Training of the neural posterior estimator was carried out on compute nodes equipped with Intel Xeon Platinum 8358 CPU nodes. We found both simulating the trajectories and performing the neural posterior estimation to be linear in the number of samples. Specifically, the total computational time to calculate the posterior with 500,000 training samples was approximately 53 minutes. Of this total runtime, the simulation of the trajectories required only 29 seconds, while the neural posterior estimation took the majority of the time with approximately 52.6 minutes. Generating 10,000 posterior samples from the trained estimator for a single time-series segment takes approximately 0.5 seconds. For the simulation of frequency trajectories, we used the implementation of *Diffrax* [16]. Further discussion of the complexity of NPE can be found in [12].

C Selection of Summary Statistics for ABC Rejection

We select these statistics as summary statistics in order to compare simulated data and observations using ABC rejection.

- Initial value
- Mean
- Standard deviation
- Variance
- Minimum
- Maximum
- Median
- 25th percentile
- 75th percentile
- Autocorrelation at lag 1
- Autocorrelation at lag 5
- Autocorrelation at lag 10
- Mean of increments
- Standard deviation of increments
- Maximum absolute increment

D Impact of Correlated Noise on Reliability of Posterior Distributions, and Limitations

In order to validate the reliability of posterior distributions under the assumption of equation (4), we estimate posterior parameter distributions for time series segments with correlated noise, and analyze the impact on the posterior estimates. Here, we simulate time series by taking samples from the prior distributions of c_1, c_2, P_0, P_1, and ϵ, and generate synthetic frequency time series with three different types of noise:

- Uncorrelated Gaussian noise ($H = 0.5$), which was used when training the neural posterior estimator
- Various degrees of temporally negatively correlated noise ($H < 0.5$), i.e. fractional Gaussian noise with a Hurst exponent of 0.2, 0.3, and 0.4
- Various degrees of temporally positively correlated noise ($H > 0.5$), i.e. fractional Gaussian noise with a Hurst exponent of 0.6, 0.7, and 0.8

For $H \neq 0.5$, the Gaussian noise in Eq. 4 is replaced by fractional Gaussian noise in order to generate time series trajectories. We can now infer posterior distributions and calculate the corresponding z-scores for these time series. Since a distribution of z-scores with an approximate mean of 0 and variance of 1 corresponds to a well-calibrated posterior, we examine the means and variances for the various time series; see Figure 8. We observe that negatively correlated noise has only a minor effect on the z-scores. Positively correlated noise, on the other hand, introduces a bias in the distribution of some of the parameters and leads to overconfidence on the part of the estimator, thereby rendering the posterior unreliable. In both cases, the estimate of the noise amplitude ϵ cannot approximate the true parameter, since the effect of the noise amplitude varies significantly depending on the type of noise correlation. The other parameters yield z-scores that are approximately normally distributed up to a Hurst exponent of 0.6, implying that the posterior distributions are well calibrated.

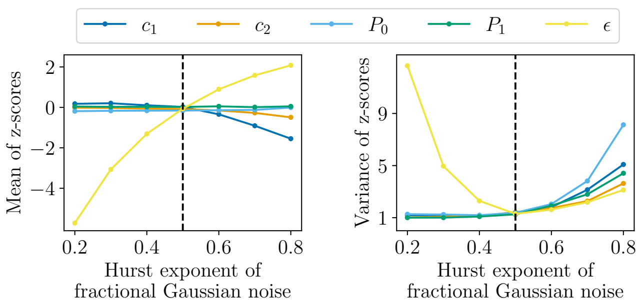

Figure 8: Influence of negatively and positively correlated noise in synthetic observations on the inference performance.

E Best and Worst Simulations for the Nordic Grid

We present, for each model, the Nordic grid trajectories corresponding to the best and worst energy scores of the NPE model in Fig. 9, complementing the CE plot from the main text. Again, we note that for the worst model (panels e-h), the assumed stochastic equation fails to capture the observed dynamics. Here, the grid displays extended oscillations, which would require a different dynamical model. Meanwhile, the NPE approach provides a very accurate and narrow estimate for its best case, clearly outperforming the partially fitting ABC and LSE baselines, and the considerably wider Daily Pattern (see panels a-d).

F Temporal Trends in Posterior Distributions in the Nordic Grid

We repeat the analysis from 5.3 for the Nordic grid, as shown in Fig. 10. Analysis of the daily profile of the various parameters reveals qualitative characteristics similar to those in the CE synchronous area. As in CE, we see strong fluctuations at the beginning of each hour, particularly in the parameters c_1, c_2, and P_0. However, we see that the fluctuations in the mean value, for example for c_1 and P_1, are significantly lower than for CE. It is also noticeable that the noise amplitude level is significantly higher at $3\,\mathrm{s}^{-3/2}$ to $4\,\mathrm{s}^{-3/2}$ than in CE. This is consistent with our observation from Fig. 5b, where the posterior predictive samples exhibit high variance in the increments, and thus in short time-scale fluctuations. With regard to the standard deviation, the values of all five parameters are in similar ranges as for CE, and the daily profile is also comparable. The width of the distributions in the daily profile is thus estimated to be similar in both networks.

G Correlation between Observation and Posterior Distribution Variance for the Nordic Grid

In the same analogous way as in 5.4, we calculate correlations of frequency time series segments and the posterior behavior for frequency data from the Nordic grid and make similar qualitative observations as for Continental Europe, see Fig. 11. It is noteworthy that, apart from the amplitude of the stochastic fluctuations ϵ, the correlations are slightly weaker than for CE. We observe a slight negative correlation with regard to the amplitude of the noise.

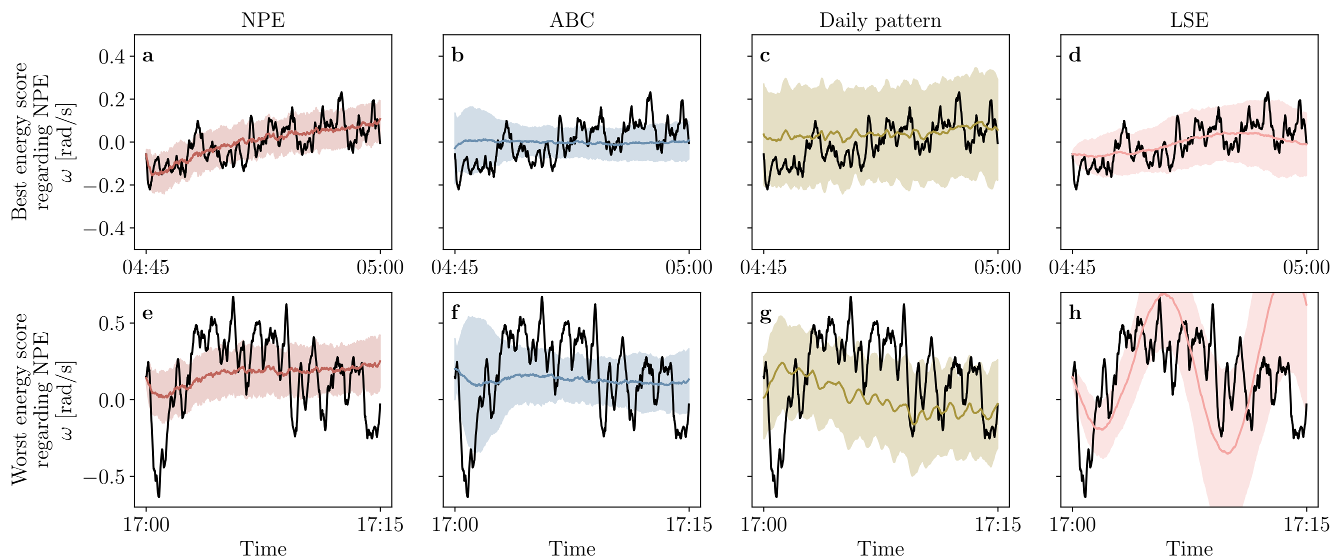

Figure 9: Comparison of observed frequency segments from the Nordic grid and trajectories simulated by 100 samples from the posterior. Black curves are the empirical frequency segments, shaded areas indicate the standard deviation of the trajectories, and colored lines represent the mean of the ensemble of trajectories. a-d: Observation with the best energy score with respect to the NPE approach, with the corresponding trajectories of all four models. e-h: Observation with the worst energy score with respect to the NPE approach, with the corresponding trajectories of all four presented models.

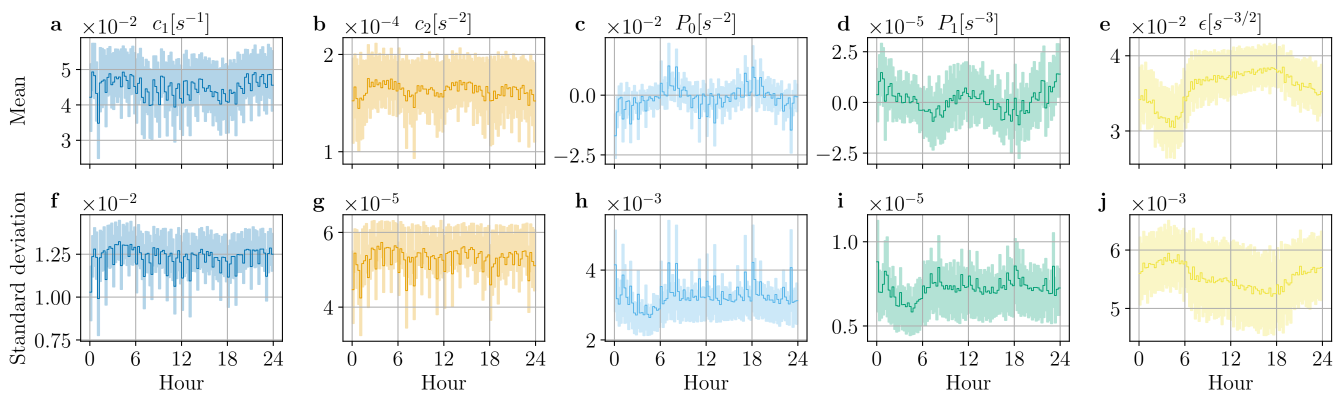

Figure 10: Daily patterns in mean (a-e) and standard deviation (f-j) of the posterior distribution of the model parameters in the Nordic grid. Solid lines represent the daily means, and shaded areas indicate the range between the 25% and 75% quantiles.

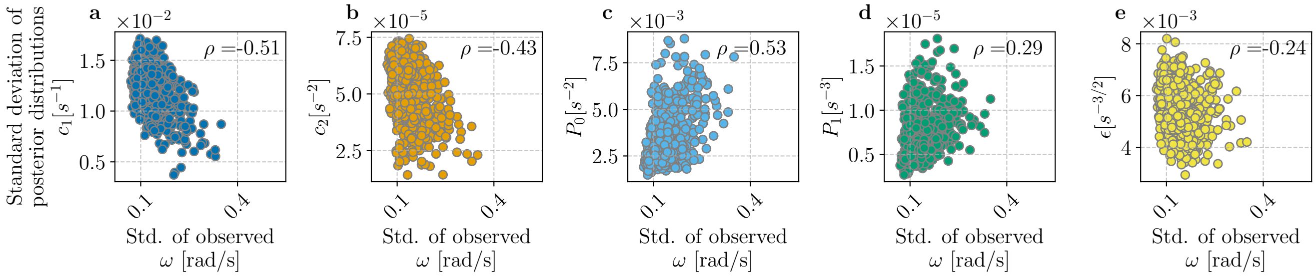

Figure 11: Relation between the standard deviation of Nordic frequency time series segments and the deviation of the corresponding parameters from the inferred posterior. The Pearson correlation coefficient between the corresponding standard deviations is given by ρ.