Sparse Variational Gaussian Processes Model for Predicting Energy Parameters in Solar Power Plants and Wind Farms

Konstantin Koshelev
ISP RAS
Moscow, Russia
e-mail: k.koshelev@ispras.ru

Sergei Strijhak
ISP RAS
Moscow, Russia
e-mail: s.strijhak@ispras.ru

Ilia Stulov
ISP RAS
Moscow, Russia
e-mail: i.stulov@ispras.ru

***Abstract*—This study applies Sparse Variational Gaussian Processes (SVGP) for probabilistic forecasting of energy output from a solar plant and a wind farm with eight turbines. Trained on two years of hourly SCADA data, the SVGP model—combining a neural mean function and kernel-based covariance—produces accurate forecasts with confidence intervals over 24–120 hour horizons. Results confirm its scalability and precision thanks to evaluation metrics such as MAE, RMSE, CRPS, and R^2 across renewable energy types.**

***Keywords*—SVGP, Gaussian Processes, probabilistic forecasting, renewable energy, solar power, wind farm, time series, metrics.**

I. Introduction

Wind and solar energy have become integral to Russia's growing sustainable energy infrastructure, with over 25 wind farms and 10 solar plants added in the past five years. These developments align with global efforts to reduce carbon emissions and contribute several gigawatts to the national grid.

For energy providers, accurate forecasting based on SCADA and meteorological data is essential to ensure stable power delivery, optimize trading, and meet regulatory requirements. Open-access web services (e.g., open-meteo.com) supply weather variables such as wind speed, direction, and solar radiation, enabling predictive models that match energy supply to demand [1], [2].

Despite advances, maintaining forecast errors within 5–10% and producing hourly predictions up to two weeks remains difficult. Standard models like ARIMA or LSTM perform well for short-term horizons (3–6 hours) but degrade over longer spans [3]. Recent research has shifted toward transformer-based methods (e.g., Chronos) and probabilistic models that better handle uncertainty [4], [5], [6]. Sparse Variational Gaussian Processes (SVGP) offer a scalable and probabilistic alternative, capable of efficient learning from large SCADA and weather datasets [7], [8], [9], [10], [11], [12], [13].

However, many existing models target individual wind turbines, whereas energy is managed and traded at the farm level, where environmental dependencies are collective. This work focuses on group-based forecasting for wind farms and solar power plants, targeting 90–95% accuracy to support real-world dispatch and energy market needs [14], [15], [16].

II. Mathematical Model

To address overfitting and underfitting in parametric models, we adopt a non-parametric Bayesian framework where the function $f(x)$ is modeled as a Gaussian Process (GP):

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')).$$

Here, $m(x)$ and $k(x, x')$ denote the mean and kernel functions, respectively:

$$m(x) = \mathbb{E}[g(x)],$$

$$k(x, x') = \mathbb{E}[(g(x) - m(x))(g(x') - m(x'))].$$

The posterior over f given data X is Gaussian:

$$p(f \mid X) = \mathcal{N}(f \mid \mu, K),$$

with observation noise modeled as:

$$p(\xi \mid f) = \mathcal{N}(\xi \mid f, \sigma^2_{\text{obs}} I).$$

SVGP Model

To handle large datasets, we use the Sparse Variational Gaussian Process (SVGP) model with inducing variables Z and function values $u = f(Z)$. This reduces complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(NM^2)$ [11], [12], [13].

The joint model is:

$$p(\xi, u \mid X, Z) = \int p(f \mid X, u) p(u \mid Z) p(\xi \mid f) \, df.$$

The variational objective maximizes the ELBO:

$$\mathcal{L}_1 = \sum_{i=1}^{N} \log \mathcal{N}(\xi_i \mid \mathbf{k}_i^\top K_{MM}^{-1} u, \sigma^2_{\text{obs}}) - \frac{1}{2\sigma^2_{\text{obs}}} \text{Tr}(\Sigma^*) + \log p(u \mid Z),$$

https://doi.org/10.51408/csit2025_63

where $\mathbf{k}_i$ is the covariance vector of inducing points, and Σ^* is the correction term.

Predictions are made via:

$$p(y \mid x) = \int p(y \mid w) p(w \mid X, Y) \, dw.$$

Predictive Mean and Variance

For a new input x_i, the predictive mean and variance are [17]:

$$\mu_f(x_i) = \mathbf{k}_i^\top K_{MM}^{-1} \mathbf{m},$$

$$\sigma_f^2(x_i) = \mathbf{k}_i^\top K_{MM}^{-1} S K_{MM}^{-1} \mathbf{k}_i + \Sigma_{i,i}^*.$$

Covariance Function (RBF Kernel)

We use the RBF kernel:

$$k(x, x') = c \cdot \exp\left(-\frac{1}{2} \sum_{d=1}^{D} b_d (x_d - x_d')^2 \right),$$

where c is the output variance, and b_d are inverse squared lengthscales.

Logit-Normal Transformation

To handle outputs $y \in (0, 1)$, we apply the logit transformation:

$$\xi = \ln\left(\frac{y}{1 - y}\right),$$

$$p(\xi \mid f) = \mathcal{N}(\xi \mid f, \sigma_{\text{obs}}^2 I).$$

III. Definition of the Problem

Forecasting energy output from renewable sources is complicated by the inherent variability of solar and wind conditions. This work applies the SVGP framework to provide scalable, probabilistic forecasts for a 10 MW solar power plant and an 8-turbine wind farm, offering both point predictions and uncertainty estimates essential for grid stability and operational planning.

A. Data Preparation

Both datasets used in this study were collected from industrial renewable energy facilities equipped with SCADA systems. All data were recorded with an hourly resolution over a two-year time span. Each dataset includes a distinct set of input features selected according to the operational characteristics of the respective energy source — solar panels or wind turbines.

B. Dataset 1: Solar Power Plant

The first dataset was collected from a 10 MW photovoltaic power station with SCADA infrastructure, covering the period from 15 March 2023 to 15 March 2025. It contains six input features: air temperature (`T_AIR`), wind speed (`V`), wind direction (`DIRECTION`), panel temperature (`T_PANELS`), horizontal irradiance (`INSOLATION_HOR`), and total insolation (`INSOLATION`). These variables were selected based on domain knowledge and prior research in solar forecasting [14]. Figure 1 illustrates the long-term behavior of total insolation.

The target variable is active power output (`POWER`), expressed in megawatts (MW). The forecasting task aims at predicting solar output over a 120-hour horizon (5 days). The SVGP model was trained using 5% of the data as inducing inputs, balancing prediction quality with training efficiency.

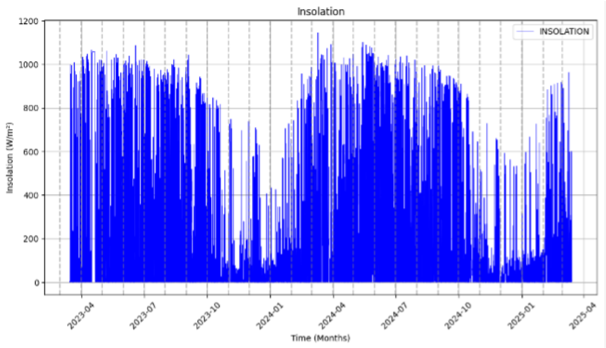

Fig. 1. Total insolation over the two-year observation period

C. Dataset 2: Wind Farm with a Group of 8 Wind Turbines

The second dataset covers a wind farm consisting of eight turbines, with data collected from SCADA systems over two years (01 January 2022 to 01 January 2024). Unlike single-turbine forecasting, this setup captures the aggregated output of the entire group, allowing the SVGP model to learn collective dynamics driven by shared atmospheric conditions. The target variable is the total power output in kilowatts (kW), with a 24-hour forecasting horizon to assess mid-term variability.

Input features include horizontal wind speed (`ABSU`), wind direction (`V10`), air temperature (`U100`), and rotor frequency (`V100`). Figure 2 shows wind speed fluctuations, illustrating the irregular and stochastic nature of wind power generation.

IV. Results

The results of this work aim to contribute to the development of scalable and accurate probabilistic forecasting systems in the energy sector. SVGP-based models provide not only point predictions but also an estimate of how uncertain those predictions are. This helps energy planners and operators better understand possible outcomes and make more informed, risk-aware decisions.

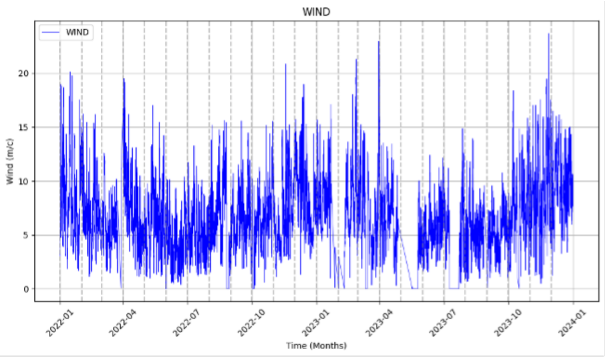

Fig. 2. Horizontal wind speed over the observation period

A. Solar Power Plant

To evaluate the SVGP model for long-term solar forecasting, we used SCADA data from a solar power plant with a 120-hour prediction horizon. The system spans 43 hectares and contains 41,184 photovoltaic panels grouped into four electrical cells. The model was trained for 200 epochs, and the loss curve (Fig. 3) shows stable convergence without signs of overfitting or underfitting.

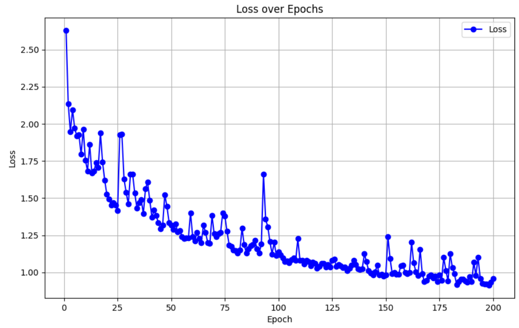

Fig. 3. Training loss over 200 epochs

1) Daily Forecasting Performance: To evaluate the stability and robustness of the SVGP model, we performed five separate 24-hour forecasts within a 120-hour test window. The corresponding prediction plots are presented in Figure 4 for the first day, showing predicted values, true observations, and 95% confidence intervals.

Table I summarizes the evaluation metrics for the five-day period and displays general performance.

2) General Forecasting Performance: Figure 5 illustrates the predicted solar power output over the 120-hour window, along with confidence intervals (95%) and true observed values. The model successfully captures the diurnal cycle of solar energy production, characterized by distinct peaks during daylight hours and near-zero output at night. The uncertainty bands widen around peak production hours, reflecting higher variability due to weather-related factors such as insolation.

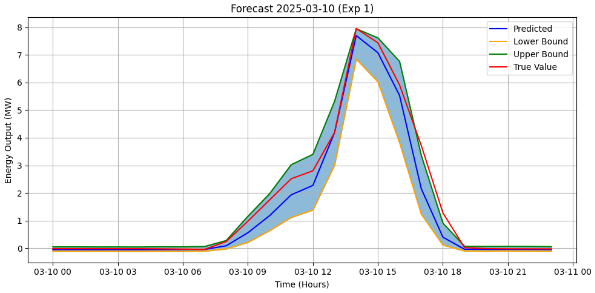

Fig. 4. 24-hour forecast for March 10, 2025

TABLE I
DAILY FORECASTING METRICS FOR SOLAR OUTPUT (SVGP)

Date	CRPS	MAE	RMSE	R^2
2025-03-10	0.023	0.031	0.055	0.967
2025-03-11	0.027	0.042	0.064	0.928
2025-03-12	0.016	0.027	0.039	0.876
2025-03-13	0.018	0.029	0.042	0.886
2025-03-14	0.018	0.024	0.034	0.972

B. Wind Farm with a Group of 8 Wind Turbines

The SVGP model was also applied to a 24-hour forecasting task for a wind farm consisting of eight wind turbines (group-1). The training was conducted over 200 epochs. Figure 6 presents the model's output. The prediction closely follows the actual curve across all 24 hours, with reasonable confidence intervals. This demonstrates the model's ability to capture middle-term wind variability.

To evaluate model performance, we computed RMSE and MAE for each individual turbine and for the combined forecast across the entire wind farm. Table II shows that individual errors remain low and consistent across turbines, with the overall aggregated RMSE reaching as low as 0.011 and MAE 0.009. These results confirm the effectiveness of SVGP in providing high-resolution, low-error predictions for wind farm energy output.

TABLE II
RMSE AND MAE FOR INDIVIDUAL TURBINES AND TOTAL WIND FARM FOR 8 WIND TURBINES PREDICTION

Turbine	RMSE	MAE
Turbine 1	0.024	0.018
Turbine 2	0.025	0.022
Turbine 3	0.034	0.028
Turbine 4	0.031	0.025
Turbine 5	0.026	0.022
Turbine 6	0.018	0.014
Turbine 7	0.018	0.014
Turbine 8	0.017	0.012
Wind farm group-1	0.011	0.009

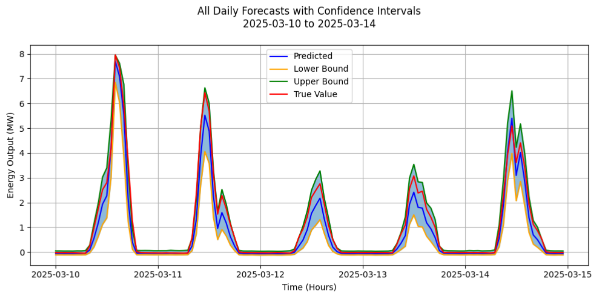

Fig. 5. Predictions with 95% confidence intervals for 120-hour solar output

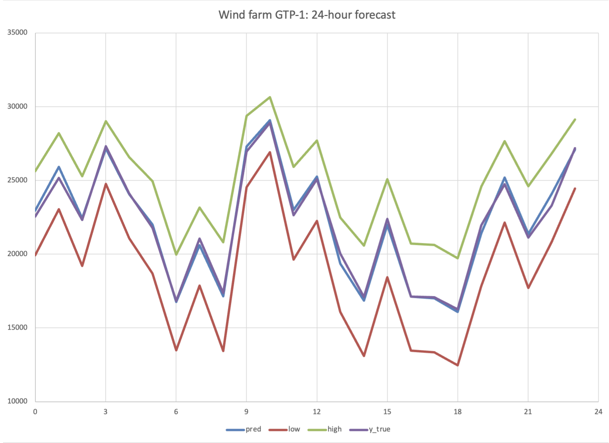

Fig. 6. 24-hour forecast for wind farm group-1 with uncertainty bounds. Axis Y: power (kW), axis X: time (hours)

V. Discussion

While the SVGP model demonstrates strong predictive accuracy and robust uncertainty quantification, several directions remain open for further research. Comparative benchmarking with other models, such as DeepAR, Temporal Fusion Transformers, Chronos [4], or support vector regression, would help evaluate trade-offs in accuracy, efficiency, and generalization on the same SCADA datasets. These models, particularly transformer-based ones may offer advantages for long-range forecasting.

Future studies could also focus on hyperparameter tuning, including the number of inducing points, kernel selection, and a neural network structure for the mean function. Additionally, incorporating time-aware feature engineering, such as sine/cosine encodings of temporal cycles (e.g., hour of day, day of year), could improve performance—especially for solar datasets with strong seasonal patterns.

VI. Conclusion

This study evaluated the performance of the SVGP model for mid- and long-term energy forecasting using historical SCADA and weather data. The model was tested on two datasets: a solar power plant with a 120-hour prediction horizon and a wind farm of eight turbines with a 24-hour horizon. In both cases, the SVGP achieved high accuracy, with RMSE not exceeding 0.03 from the maximum power.

For the solar dataset, five daily forecasts were generated, each maintaining a high R^2 (above 87%), with better performance on days with higher insolation. Feature analysis indicated that insolation-related variables had the strongest influence. While no features were excluded in this study, adding more input parameters or extending the dataset could improve generalization.

Model training under a 5% inducing point configuration (about 900 records) took less than 10 minutes, highlighting its efficiency. The use of historical weather data also suggests future extensions with probabilistic forecasts for enhanced planning and operational utility.

References

[1] T. Wen-Chang, H. Chih-Ming, T. Chia-Sheng, L. Whei-Min, and C. Chiung-Hsing, “A Review of Modern Wind Power Generation Forecasting Technologies,” *Sustainability*, vol. 15, no. 14, p. 10757, 2023.

[2] Y. Zhang, J. Wang, and X. Wang, “Review on probabilistic forecasting of wind power generation,” *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 255–270, 2014.

[3] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, “Energy Forecasting: A Review and Outlook,” *IEEE Open Access Journal of Power and Energy*, vol. 7, pp. 376–388, 2020.

[4] A. F. Ansari *et al.*, “Chronos: Learning the language of time series,” *Transactions on Machine Learning Research*, 2024.

[5] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, “GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration,”, 2021. [Online]. Available: https://arxiv.org/abs/2104.13628.

[6] A. Potapczynski, M. Finzi, G. Pleiss, and A. G. Wilson, “CoLA: Exploiting Compositional Structure for Automatic and Efficient Numerical Linear Algebra,”, 2023. [Online]. Available: https://arxiv.org/abs/2309.03060.

[7] K. A. Wang, G. Pleiss, J. R. Gardner, S. Tyree, K. Q. Weinberger, and A. G. Wilson, “Exact Gaussian Processes on a Million Data Points,”, 2019. [Online]. Available: https://arxiv.org/abs/1903.08114.

[8] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*, MIT Press, 2022. [Online]. Available: https://books.google.ru/books?id=HLIyzgEACAAJ

[9] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*, MIT Press, 2023.

[10] C. E. Rasmussen and H. Nickisch, “Gaussian processes for machine learning (GPML) toolbox,” *J. Mach. Learn. Res.*, vol. 11, pp. 3011–3015, 2010.

[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[12] J. Hensman, N. Fusi, and N. D. Lawrence, “Gaussian processes for big data,”, 2013. [Online]. Available: https://arxiv.org/abs/1309.6835.

[13] J. Hensman, A. Matthews, Z. Ghahramani, “Scalable Variational Gaussian Process Classification,” In: *Proc. of the Eighteenth Int. Conf. on Artificial Intelligence and Statistics*, San Diego, California, USA, PMLR, pp. 351–360, 2015.

[14] E. Zelikman, S. Zhou, J. Irvin, C. Raterink, H. Sheng, A. Avati, J. Kelly, R. Rajagopal, A. Y. Ng, and D. Gagne, “Short-Term Solar Irradiance Forecasting Using Calibrated Probabilistic Models,” arXiv preprint arXiv:2010.04715v2, 2020. [Online]. Available: https://arxiv.org/abs/2010.04715

[15] K. Doubleday, S. Jascourt, W. Kleiber, and B. M. Hodge, “Probabilistic solar power forecasting using Bayesian model averaging,” *IEEE Trans. Sustain. Energy*, vol. 12, pp. 325–337, 2021.

[16] A. K. Tripathi *et al.*, “Advancing solar PV panel power prediction: A comparative machine learning approach in fluctuating environmental conditions,” *Case Stud. Therm. Eng.*, vol. 59, p. 104459, 2024.

[17] H. Wen, J. Ma, J. Gu, L. Yuan, and Z. Jin, “Sparse Variational Gaussian Process Based Day-Ahead Probabilistic Wind Power Forecasting,” *IEEE Trans. on Sustainable Energy*, vol. 13, pp. 957–970, 2022.