Impact of Climate and Energy Consumption on Industrial Activity Index: A Regression Analysis

Iclal Cetin Tas[1]

***Abstract*—**Industrial activities and emissions are affected by climate change and energy use. The system that models industrial indicators is very complicated and can change a lot depending on what is happening outside. This makes it hard to use only climate and energy variables to explain how the system worksThis study examines the explanatory capacity of industrial activity using climate and energy data from 21 countries for five years, employing heteroskedastic Gaussian Process Regression (GPR) and Automated Attribute Determination (ARD) based feature selection methods. The findings demonstrate that climate and energy variables have a limited but consistent explanatory effect on industrial activity. All experiments conducted showed that the MASE values obtained were lower than 0.50. The proposed method surpassed other state-of-the-art methods, achieving a MASE value of 0.44.The study found that the ARD-GPR method, which considers how uncertain the data is more accurate than other methods in predicting how much energy and climate change affect industrial activity.

***Index Terms*—** Climate Change, Energy, Industrial Activity Index, Regression.

I. Introduction

The industrial index is a composite measure derived from integrating various multidimensional economic indicators related to the industrial sector through statistical methods. Its primary purpose is to encapsulate short- and medium-term fluctuations in industrial production, changes in production capacity, and production dynamics associated with energy consumption into a single quantitative metric. In this context, the industrial activity index is frequently utilized as a dependent variable in analyzing relationships with external factors such as energy use, climatic conditions, and macroeconomic variables. [1][2].

Energy consumption in the industrial sector is critical for economic growth, sustainability, and competitiveness. Accurately forecasting energy consumption in industrial facilities provides multifaceted benefits such as production planning, cost control, resource efficiency, and reduction of environmental impacts. Electricity is a fundamental input in industrial production and is indispensable for the continuity and quality of production processes. Industrial activities are indirectly affected by climate conditions as well as energy consumption indicators. While climate variables such as average temperature and humidity play a decisive role in energy demand and production processes, energy and socio-economic factors such as carbon dioxide emissions, renewable energy share, energy prices, and urban population ratio are among the key elements shaping the level of industrial activity. Therefore, regression-based analyses that consider multidimensional variables related to climate and energy together enable a more realistic and holistic modelling of the industrial activity index.

Eke's study [3] used a combination of differential evolution algorithms and artificial neural networks. The study found that the hybrid model, which used both DE and artificial neural networks, was better than traditional artificial neural network models. A new model for predicting how much energy a factory will use in the next few months was made. It uses advanced methods to find the best way to do this. The model shows that these methods work well for this kind of prediction.

Some models that use Gaussian Process Regression (GPR) can predict how much load a structure can handle and how much uncertainty there is in the prediction. They can do this because they use a mathematical formula that combines different parts of the data. GPR-based methods can capture how much load a structure can handle in different situations and how well they can handle different situations. They can also handle more situations than classical deterministic models [4]. GPR can make the predictions more reliable and precise for medium and weekly load estimates. It can also do better than other methods that use patterns from the past, such as ARIMA [4], for the same purpose. The study uses a method that uses GPR to estimate the load on a structure. GPR can handle uncertainty and change easily.

In short-term electricity load forecasting problems, models capable of addressing both temporal dependencies and uncertainty structures are needed due to high variability and sudden load fluctuations. In this context, approaches that consider GPR as the fundamental probabilistic inference mechanism and use it in hybridization with models that have stronger representation capabilities in short-term forecasting stand out in the literature. In particular, the BiGRU-GAM-GPR-based study integrated GPR with deep learning-based time series models by positioning it at the center of short-term load forecasting, and demonstrated that this hybrid structure increases both forecasting accuracy and probabilistic reliability [6].

GPR's flexible core-based structure and inherent uncertainty modelling capability allow for the efficient representation of nonlinear dynamics occurring at different time scales. This model is suitable for solution of r both short-term and long-term forecasting problems. GPR can be further enhanced by

[1]Department of Computer Engineering, Baskent University, Turkiye

hybridizing it with models that have high temporal learning capacity for short-term forecasting, and with methods that can capture trend and structural information for long-term forecasting. GPR can offer more comprehensive and reliable forecasting frameworks when integrated with methods possessing different capabilities.

The study continues with the dataset information the preprocessing steps, and the process of creating the industrial activity index in Materials and Methods section. It also presents the mathematical basis of the GPR framework, which includes a linear mean function, an ARD-based kernel structure, and a heteroscedastic noise model. The Results and Discussion section evaluates and discusses the prediction performance of the proposed model using different error metrics. Conclusion section summarizes the findings and evaluates future research directions.

II. Materials and Methods

A. Materials and Methods

This study uses the publicly dataset for all experiments. Climate and Energy Consumption Dataset published on Kaggle [7], to analyze the relationship between energy consumption and climate variables. The dataset covers the years 2020–2024 and includes climate and energy consumption data collected on a daily scale. The dataset contains meteorological parameters such as temperature, humidity, wind speed, and precipitation, as well as variables related to total energy consumption. In this study, the dataset was normalized and divided into training and test sets for use in the analysis and modelling phases. The ratio is %80 train, %20 test for this study.

B. Methods

This study uses seven different models encompassing parametric, regularized, nonlinear, and probabilistic regression approaches to model energy consumption in relation to climate variables. This diversity aims to comparatively evaluate different modelling assumptions and generalization capabilities.

The linear regression model is a simple way to compare how well the dependent variable depends on the independent variable. It assumes that the relationship is straight and constant. The model is based on the classical linear assumption that parameters are estimated using the least squares approach.

Lasso Regression is a linear regression method that includes regularization based on the L1 norm. - This method makes the model simpler and faster by getting rid of some features that do not matter much. - It also helps to choose the best features by making the less important ones more negative.

Elastic Net Regression is a hybrid approach that uses both L1 and L2 regularizations. This way of doing things gives more fair and accurate numbers, especially when the variables are related to each other, like weather data.

Polynomial Regression is a type of linear regression that can model nonlinear relationships by using different powers of independent variables. We can use this method to create models that show how different climate factors affect how much energy people use.

The Naive (Persistence) model is a basic way of estimating how much energy is used by a system. It uses the energy value from the last step as a reference. This model has a lower limit, especially for problems that involve time series data, to measure how well other methods work.

TABLE I: State of Arts (SOTA) Regression Models Mathematical Explanation

Model	Mathematical Formulation	
Linear Regression	$\hat{y} = \beta^0 + \Sigma_i^{=1n} \beta_i x_i$	(1)
Lasso Regression	$\hat{\beta} = \arg\min[\Sigma_i^{=1N} (y_i - x_i^T\beta)^2 + \lambda \Sigma\square^{=1p} \lvert\beta\square\rvert]$	(2)
Elastic Net Regression	$\hat{\beta} = \arg\min[\Sigma_i^{=1N} (y_i - x_i^T\beta)^2 + \lambda^1\Sigma\square^{=1p} \lvert\beta\square\rvert + \lambda^2\Sigma\square^{=1p} \beta\square^2]$	(3)
Polynomial Regression	$\hat{y} = \beta^0 + \beta^1 x + \beta^2 x^2 + \ldots + \beta_d x^d$	(4)
Naive (Persistence) Model	$\hat{y}\square = y\square^{-1}$	(5)

The Heteroscedastic GPR model is a probabilistic, nonparametric regression approach. The variance of the output variable is allowed to vary as a function of the input features, rather than being assumed constant. This formulation is particularly suitable for modelling energy consumption and industrial activity, where noise levels and volatility change over time due to external influences such as climatic variability, energy price fluctuations, and structural economic dynamics. By explicitly modelling input-dependent noise, heteroscedastic GPR can capture complex uncertainty patterns and variable noise structures within the data, providing not only point predictions but also associated uncertainty estimates that reflect the confidence of each forecast.

$$f(x) \sim \mathcal{GP}\big(\mu(x), k_{ARD}(x, x')\big) \quad (6)$$

$$\varepsilon(x) \sim \mathcal{N}\big(0, \sigma^2(x)\big) \quad (7)$$

$$\mu(x) = \beta_0 + \sum_{d=1}^{D} \beta_d x_d \quad (8)$$

$$k_{ARD}(x, x') = \sigma_f^2 exp(-\frac{1}{2}\sum_{d=1}^{D} \frac{x_d - x'_d}{l_d^2})) \quad (9)$$

$$Cov(y_i, y_j) = k_{ARD}(x_i, x_j) + \sigma^2(x_i)\delta_{ij} \quad (10)$$

The proposed model employs a heteroscedastic GPR framework with a linear mean function and an ARD-based kernel, where the observation noise variance is modelled as an input-dependent function to capture non-stationary uncertainty across the input space.

To address scalability and enhance predictive performance, the heteroscedastic GPR model is integrated into a hybrid framework with a linear regression component. In this hybrid approach, the linear model is first used to capture the dominant linear trends and global relationships between climate and energy indicators and the industrial index. The heteroscedastic GPR is then applied to the residuals of the linear model, allowing it to focus on nonlinear dependencies and input-dependent variance that cannot be explained by linear assumptions alone. This residual-based hybridization reduces the complexity of the Gaussian Process component, improves computational efficiency on large datasets, and enables more accurate modelling of heteroscedastic effects, ultimately resulting in improved error performance and more reliable uncertainty-aware forecasts

III. RESULTS AND DISCUSSION

This study addresses the estimation of an industrial index representing industrial activity using climate and energy indicators. The industrial index exhibits heteroskedastic behaviour, with both its mean level and variability changing over time, indicating a non-constant error variance. Therefore, models based on the assumption of constant error variance have limited ability to fully represent the dynamic nature of industrial activity. For this reason, five different SOTA methods focusing on constants and variables were used in the analyses. In addition, the heteroscedastic method linear GPR, known to be successful in uncertainty analysis, was used in hybridization. All results obtained are shown in Table II.

TABLE II: Obtained Results

Model	RMSE	MAE	MASE
Linear Regression	0.5684	0.4596	0.4638
Polynomial Regression (deg 2)	0.5686	0.4598	0.4640
Lasso Regression	0.5694	0.4603	0.4645
Elastic Net Regression	0.5701	0.4609	0.4651
Naive (Persistence) Model	1.3686	1.2025	1.2135
Hybrid: Linear + Heteroscedastic GPR	0.5493	0.4440	0.4480

ARD principle was applied to linear and GPR based models to automatically determine the relative effects of climate and energy variables on the industrial index. The linear regression model achieved RMSE = 0.568 and MAE = 0.460, while the Lasso and Elastic Net models remained at similar error levels (RMSE ≈ 0.569–0.570, MAE ≈ 0.460–0.461). These results show that while a certain portion of the industrial index can be explained by linear relationships, the variable error structure cannot be directly represented by these models. In contrast, the proposed method offered a more flexible modelling framework that could indirectly capture the heteroscedastic structure. The proposed method achieved a significant improvement in RMSE and MAE metrics compared to linear models (RMSE = 0.549, MAE = 0.444). This reduction demonstrates that the model not only reduced the average prediction error but also more effectively suppressed large deviations that occur during periods of high volatility in industrial activity.

Proposed approach also excelled in terms of the MASE metric, a scale-independent performance indicator. The proposed model achieved a MASE value of 0.448, while the MASE values of the linear and regularized models remained at approximately 0.464. This situation demonstrates that the proposed method learns the temporal dependencies and variable error structure of the industrial index more effectively compared to the naive reference model. The high error values of the naive (persistence) model (RMSE = 1.369, MAE = 1.203, MASE = 1.214) clearly reveal that industrial activity has a complex structure that cannot be explained by simple historical values.When dealing with a heteroscedastic target variable with high scale variability, scale-independent metrics such as RMSE and MASE appear to offer a more reliable assessment. When the numerical results are evaluated together, it is observed that the relationship between climate and energy variables and the industrial index has a heteroscedastic structure containing both nonlinear and time-dependent variable error variance. It has been shown that this structure can be represented more stably and with lower error levels by the GPR-based hybrid model supported by the ARD mechanism. Figure 1 shows that the proposed model successfully tracks log-difference values. Figure 2 shows that the actual and predicted series overlap and that it can capture the direction of fluctuations. While the histogram graph indicates an approximately normal distribution, the residual-predicted scatter plot shows a random distribution rather than a strong linear structure, indicating that the model is not biased. The cumulative absolute error curve reveals that the Heteroscedastic GPR model exhibits a more controlled accumulation of errors compared to other regression methods. When all the results obtained are evaluated, it is seen that the proposed approach offers statistically consistent and reliable performance in terms of both accuracy and error distribution

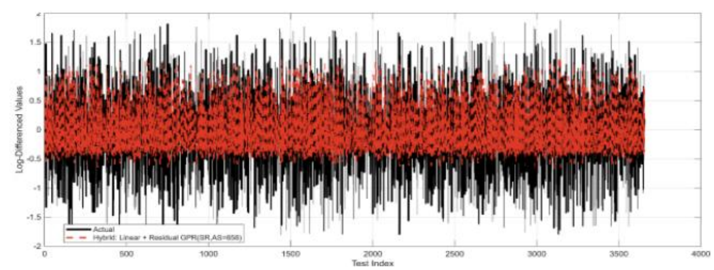

Fig. 1: Actual predicted results log-difference values

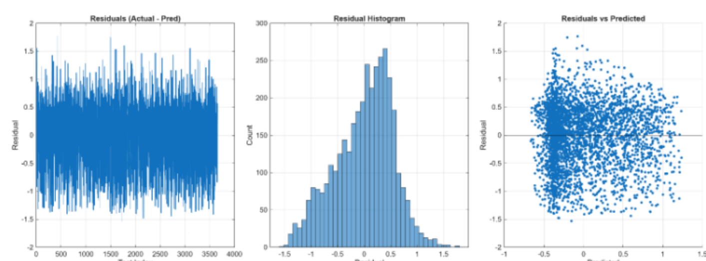

Fig. 2: Residual predicted results comparison

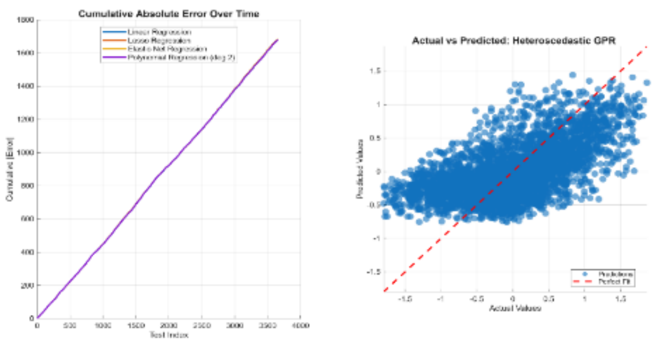

Fig. 3: a) Cumulative Error b)Actual-Predicted Values

IV. CONCLUSION

This study addresses the problem of modelling the industrial index using climate and energy indicators which compares the ARD principle within the framework of linear and GPR-based approaches. The aim is to explain the extent to which industrial activity overlaps with climate-energy interactions. The linear model can show how the industrial index changes over time in a simple way, but it cannot make the changes smaller when there are complex and nonlinear effects. The hybrid approach that combines linear ARD and GPR methods gave better and more equal results in basic error measures like RMSE, MAE and MASE. The study shows that the proposed method, which takes into account how uncertain the data is, is more suitable for predicting how much energy and how much climate change is caused by industrial activity than the usual methods that assume the data is always the same and follow a straight line. GPR-based hybrid models that can change according to the data size and type are helpful for saving time and avoiding errors, especially when working with large and varied data.

Future studies plan to model the industrial index separately through its sectoral and relational subcomponents and to conduct comparative evaluations for different country groups. This will contribute to a more detailed analysis of the impacts of climate and energy policies on industrial activity.

REFERENCES

[1] Y. Zhao, R. Yang, Y. Jie, S. Wang, and Z. Zeng, "Enhancing online industrial quality index prediction with a general deep temporal feature extraction and incremental ensemble modeling framework," *Engineering Applications of Artificial Intelligence*, vol. 156, p. 111104, Sep. 2025, doi: 10.1016/j.engappai.2025.111104.

[2] S. C. Shongwe, Z. Greyling, T. Zwane, and S. Gcanga, "Analyzing the South African Industrial Index Data using Numerous Standard Statistical Distributions," *Journal of Statistics Applications & Probability*, vol. 14, no. 1, pp. 27–44, Jan. 2025, doi: 10.18576/jsap/140103.

[3] I. Eke, "Medium-Term Load Forecasting by Artificial Neural Network Based Differential Evolution," International Journal of Engineering Research, vol. 3, no. 1, pp. 28–32, Jan. 2011.

[4] A. Yadav, R. Bareth, M. Kochar, M. Pazoki, and R. A. El Sehiemy, "Gaussian process regression-based load forecasting model," IET Generation, Transmission & Distribution, vol. 18, no. 5, pp. 899–910, Mar. 2024, doi: 10.1049/gtd2.12926.

[5] T. Ma, R. Huang, D. Barajas-Solano, R. Tipireddy, and A. M. Tartakovsky, "Electric Load and Power Forecasting Using Ensemble Gaussian Process Regression," arXiv preprint arXiv:1910.03783, Oct. 2019.

[6] Q. Shao, R. Bao, S. Liu, K. Fu, L. Mo, and W. Xiao, "Short-Term Electric Load Probability Forecasting Based on the BiGRU-GAM-GPR Model," Sustainability, vol. 17, no. 12, Art. no. 5267, 2025, doi: 10.3390/su17125267.

[7] E. Akkuş, "Climate & Energy Consumption Dataset 2020–2024," Kaggle Dataset, 2025. [Online]. Available: https://www.kaggle.com/datasets/emirhanakku/climate-and-energy-consumption-dataset-20202024. [Accessed: Jan. 29, 2026]..