DÜZCE ÜNİVERSİTESİ
BİLİM ve TEKNOLOJİ DERGİSİ
Düzce University Journal of Science & Technology

Düzce University
Journal of Science & Technology

Research Article

Confidence Interval Approach to Weather Forecasting with Horizon-Based Genetic Programming

Ömer MİNTEMUR

Department of Software Engineering, Ankara Yıldırım Beyazıt University, Ankara, Türkiye
Corresponding author's e-mail address: omermintemur@aybu.edu.tr

DOI: https://doi.org/10.29130/dubited.1188691

ABSTRACT

Being able to forecast events has always been important for humans. Humans performed forecasting by observing the movements of material and non-material objects in ancient times. However, with technological developments and the increasing availability of data in recent years, forecasting has been done by computers, especially by machine learning methods. One of the areas where these methods are used frequently is numerical weather forecasting. In this type of forecast, short term, medium term and long term numerical weather forecasts are made using historical information. However, predictions are inherently error-prone, and should be indicated within which error ranges the predictions fall. In this study, numerical weather forecasting was done by combining Genetic Programming and the Inductive Conformal Prediction method. The effects of 10 and 20 days of historical data on short (1 day), medium (3 days) and long term (5 days) weather forecasts were examined. The results suggested that Genetic Programming has a good potential to be used in this area. However, when Genetic Programming was combined with the Inductive Conformal Prediction method, it was shown that forecasts gave meaningful results only in the short term; forecasts that were made for the medium and the long term did not produce meaningful results.

Keywords: *Artificial Intelligence, Confidence Interval, Prediction, Genetic Programming*

Ufuk Amaçlı Genetik Programlama ile Hava Durumu Tahminine Güven Aralıklı Yaklaşım

ÖZ

Olayları önceden tahmin edebilmek insanlar için her zaman önemli olmuştur. Eski zamanlarda insanlar tahminlerini maddesel ve maddesel olmayan cisimlerin hareketlerine göre yapmışlardır. Ancak, son yıllardaki teknolojik gelişmeler ve veri miktarındaki artış sayesinde tahmin çalışmaları bilgisayarlar tarafından, özellikle de makine öğrenmesi metotları tarafından yapılmaktadır. Bu metotların kullanıldığı en önemli alanlardan bir tanesi de sayısal hava durumu tahminidir. Bu tahmin çeşidinde, tarihsel veriler kullanılarak kısa, orta ve uzun vadeli tahminler yapılmaktadır. Ancak, tahminler doğası gereği hataya açık olaylardır ve ortaya çıkan hatanın hangi aralıkta olduğu belirtilmelidir. Bu çalışmada sayısal hava durumu tahmini Genetik Programlama ve Tümevarımsal Güven Aralığı metodu birleştirilerek yapılmıştır. 10 ve 20 günlük tarihsel verinin kısa (1 gün), orta (3 gün) ve uzun (5 gün) vadede yapılan tahminlere olan etkisi incelenmiştir. Sonuçlar Genetik Programlamanın bu alanda kullanılabileceğini göstermektedir. Ancak, Genetik Programlama Tümevarımsal Güven Aralığı

Received: 14/10/2022, Revised: 08/02/2023, Accepted: 24/03/2023

metodu ile kullanılınca, yapılan tahminlerin sadece kısa vadede anlamlı sonuçlar ürettiği görülmüştür. Orta ve uzun vadeli yapılan tahminlerin anlamlı sonuçlar üretmediği görülmüştür.

Anahtar Kelimeler: *Yapay Zeka, Güven Aralığı, Tahmin, Genetik Programlama*

I. INTRODUCTION

Humans have a strong tendency to see a cosmos in nature. This tendency has encouraged humans to seek a cause-and-effect relationship in nature. As a result, forecasting mechanisms in humans have emerged. In ancient times, due to the absence of necessary technology, people made predictions by observing the behaviors of material and non-material objects, such as planets, waves, and winds. However, nowadays, humans have transferred their forecasting ability to technology. Forecasts which had been made by humans in the non-technological era have been replaced by computers thanks to development in technology and the huge amount of data that are available today. Today, forecasts are generally made by algorithms that have been developed in the area called Artificial Intelligence (AI). These algorithms try to find relations between variables in which humans may not see clearly or do not have a mathematical function to describe the relation completely. AI algorithms use historical data of a particular phenomenon to model the forecasting mechanism, with the assumption that the data contains enough information which enable AI model to make future predictions.

Forecasting methods in AI realm can be divided into two subcategories according to mechanisms of the algorithms employ. Methods such as ARIMA [1] and Prophet [2] are specialized methods used only in forecasting. Another subarea of this field consists of Deep Learning (DL) methods, such as Long Short Time Memory (LSTM) [3], Artificial Neural Networks (ANN)[4], and Machine Learning (ML) algorithms such as Support Vector Machines (SVM)[5], K-Nearest Neighbors (KNN)[6]. While DL and ML methods are more commonly used for classification, image recognition, and image generation problems, these techniques can also be applied to forecasting by manipulating them. These methods have been shown to be effective through their ability to be readily adapted and their extensive application in forecasting domains, including electricity load forecasting [7,8], stock market prediction [9], and storm forecasting [10].

Inevitably, the power of these mentioned methods has been used in one of the hardest problems in forecasting; that is, numerical weather prediction from historical data. It is known that the weather itself is a chaotic system and making numerical weather prediction is a challenging problem [11,12]. However, both DL and ML methods have been used in numerical weather forecasting. A detailed comparison of ML and DL methods can be found in [13] where the authors thoroughly analyzed the performance of these methods in forecasting the weather using the NCDC data set. Another study that suggests a DL method specifically showed that a small error rate can be achieved when predicting wind speed in the weather [14]. Since long term weather forecasting is likely prone to errors, short term forecasting is the best option for the task at hand, and such a study can be found in [15] where the authors proposed a model that improved forecast accuracy up to 3 days in advance. The weather forecasting studies involve not only temperature forecasting, but also involve forecasting the general results of weather behavior such as tsunamis, typhoons, etc. Such a study can be found in [16] where the authors proposed a DL mechanism to predict the intensity of the typhoons in advance.

However, both the DL and ML methods have several drawbacks. A major drawback of these algorithms is their parameters are determined through trial and error. As a result, each combination of parameters may yield different results. In other words, finding the parameters that give the best accuracy is an exhausting process. Besides, both DL and many of the ML methods are black-box algorithms, which means that they are not easily interpreted by humans.

The problem of these black-box approaches can be solved with Genetic Programming (GP) [17]. Due to its simplicity and power, it is also used in forecasting problems [18,19]. GP is also utilized for weather-related problems. A recent study using GP’s power predicted rainfall prediction [20]. The results indicated by the authors suggested that GP outperformed most of the fundamental ML algorithms. Another study that employed GP for water level forecasting can be found in [21].

These algorithms have proven their strengths in the forecast area by achieving high accuracies. However, especially the weather forecasting has an inevitable consequence, which is the error produced by the constructed model. It occurs especially when forecasting distant horizons [22]. This error is a numerical quantity, and it does give only bare information about the developed model. However, supporting this error quantity with a confidence interval is important in the forecasting studies. Given that weather dynamics are influenced by various independent attributes (pressure, humidity, etc.), even a slight change in one of these attributes can significantly impact the forecasting model. Thus, providing a confidence level for the model when making numerical weather predictions enhances our understanding of the model's performance.

However, forecasting methods combined with confidence interval (CI) in the literature are scarce and should be used more often. Another issue in this area is the lack of an experimental study that extensively analyzes the effect of historical data on different forecast horizons.

With these progresses and shortcomings in the literature in mind, the major problems of numerical weather predictions can be summarized as follows:

- Models that are understandable by humans are needed,
- CI-based weather forecasting should be a standard approach,
- Effects of historical data on weather forecasting should be analyzed.

In this study, GP combined with Inductive Conformal Prediction (ICP), was analyzed. Also, the effects of historical data on different forecast horizons were shown experimentally. In the present work, 10 and 20 days were used as historical information. Forecast horizons were selected as 1 day, 3 days, and 5 days. Different GP models were constructed for each horizon. Moreover, the generated models by GP were tested monthly to see their genuine performances. Dataset used in the experiments was chosen as Jena Weather Dataset which has historical weather data from 2004 to 2020 [23].

The rest of the paper is organized as follows: Section II describes the dataset, GP, ICP, and design of the experiments. Section III presents the results of the experiments. Inferences and comments about the experimental results are also given in Section III. Finally, the conclusion and possible future works are given in Section IV.

II. MATERIAL AND METHOD

In this section, the dataset used in the experiments is explored, decent information about GP and ICP are given to reader and finally the experiment schema is explained.

A. DATASET

The dataset used in the experiments is an open-source Jena dataset that possesses information such as temperature, CO2 level, pressure, etc. that was collected from the city of Jena in Germany. Sensors collected distinctive attributes of the weather at 10-minutes intervals in that region.

The aim of the paper is to forecast the daily temperature at different horizons. Thus, by taking 24-hour averages of each day, the overall dataset was converted into a format as if the data acquisition interval was utilized at daily. Since numerical weather prediction is the experimental aim of this study, the dependent variable that GP models try to forecast by minimizing the error was selected as T(degC). Other attributes in the dataset were given to GP models as independent variables. (see [23] for more information on the dataset).

Month-based distribution of T(degC) is given in Figure 1. An examination of Figure 1 indicates that there is a predictable pattern in the data for each month, due to the fact that the data were collected from a relatively small region. Another feature of the dataset is that the maximum and the minimum values of each month oscillate between certain intervals, which means that the dataset does not have extreme outliers. One more observation can be made about the stationary status of the dataset. Month-based observations reveal that each month's weather behavior repeats itself through the years. Thus, it can be concluded that the dataset is stationary, so manipulation of the dataset for stationary and seasonality is not needed. The next section briefly explains GP and ICP to readers so that they have decent information about them.

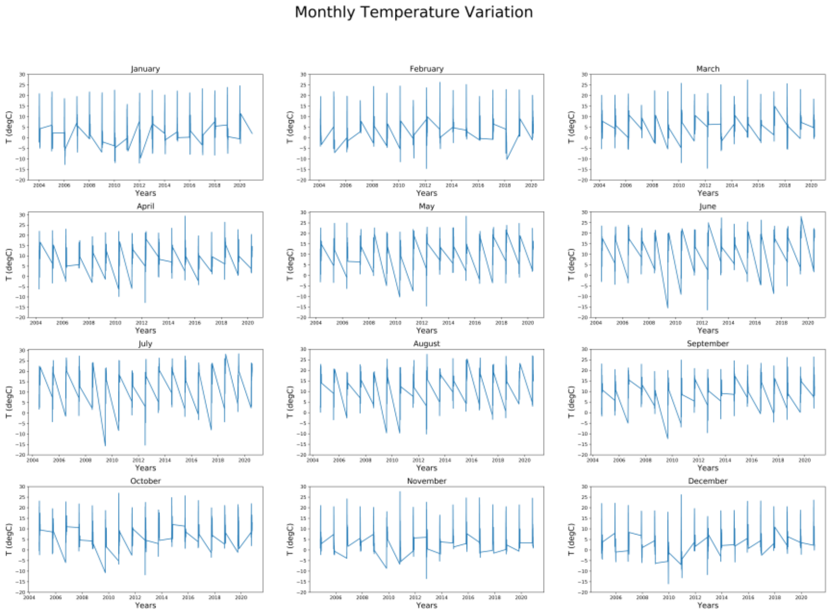
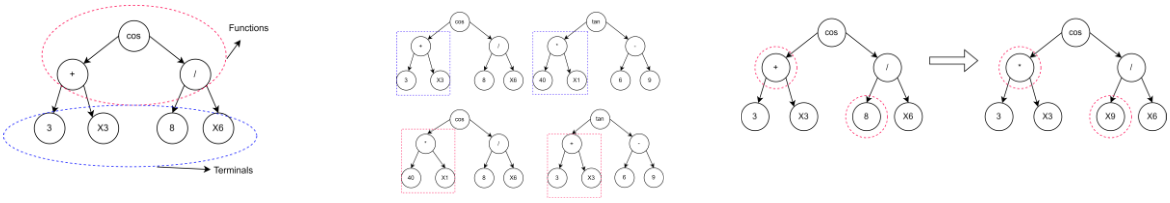


Figure 1. *Monthly temperature variation*

B. GENETIC PROGRAMMING

The Genetic Programming (GP) scheme is inspired by Charles Darwin's evolution theory. The theory indicates that the best suitable organism in nature has the highest chance of living and breeding. GP takes this idea and uses it to solve complex problems. The main concept in GP is to establish a mathematical relationship between inputs and outputs using basic mathematical functions such as addition, subtraction, and multiplication, or user-defined functions. GP begins its operations by constructing trees (called population); each tree is called an individual, and it is a possible solution to the problem. Each tree in GP has two essential components: a combination of functions selected from the function set, and a terminal set that corresponds to the attributes in the dataset. For the second step, each solution is evaluated by a fitness function that measures the error between the output that is generated by the individual and the actual output. After each solution is evaluated, the solutions generated by each individual are sorted in an ascending order (from the smallest error to the biggest error). Then, the best candidates for solutions are transferred to the next generation. To create diversity in the solutions, a crossover mechanism that combines different parts of two trees to originate a new tree and a mutation mechanism that randomly changes a function or a variable in a tree are applied. A general working mechanism of GP is shown in Figure 2.



(a) ***(b)*** ***(c)***

Figure 2. *General Working Mechanism of Genetic Programming;* ***(a)*** *A constructed tree by GP;* ***(b)*** *Crossover between two trees;* ***(c)*** *A mutation occurred in a tree*

Figure 2 (a) shows a possible solution generated by GP. X_n (X3 and X6) corresponds to attributes in the dataset. Figure 2 (b) graphically depicts the crossover mechanism between two trees, where random sub-branches of each tree are exchanged. Finally, Figure 2 (c) presents the mutation step of GP, where random changes occur in any node. The changes are made based on whether it is a function node or a terminal node.

This study uses GP since it does not require any human intervention in contrast to DL or ML methods. Another important advantage of GP is that it has a modifiable structure in terms of functions. Adding new functions is relatively easy and these functions may have a positive effect on the performance of the generated models. Finally, models that are constructed by GP are easily understandable by humans since they will be simple mathematical functions.

The next section briefly explains the ICP that was selected as the CI method in this paper.

C. INDUCTIVE CONFORMAL PREDICTION

The Inductive Conformal Prediction (ICP) is a model-free method to ensure confidence and coverage [24]. Generally, in either DL models or regression models in ML, the output is a single continuous number. However, the range of outcomes should be specified for a particular input when making a prediction using a model. Since ICP is easy to use with any algorithm in AI, it was selected in this study.

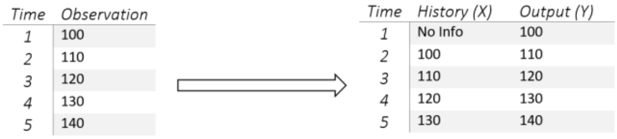
Successfully applying ICP to a forecasting model requires dividing the dataset into three distinct subsets: the train set, test set, and validation set. The overall algorithm for ICP is given below:

1. Divide the dataset into three subsets. $D_{all} = D_{train} + D_{test} + D_{val}$
2. Train D_{train} with the model M using $(x_{i,}y_i)$ set and obtain the loss L, where $L = F(x_i y_i)$ and $x_i, y_i \in D_{train}$
3. Compute L_i for each (x_i, y_i) where $(x_i, y_i) \in D_{test}$
4. Calculate quantile q using L_i from the Step 3 and $\frac{[(n_{test}+1)(1-\alpha)]}{n_{test}}$ where n_{test} is the number of points in D_{test} and α is the error rate between $(0,1)$.
5. Finally, calculate the confidence interval of the model using $[M(x_i) - q, M(x_i) + q]$ where $(x_i, y_i) \in D_{val}$.

α is selected as 0.1 in the experiments, which means that we wanted the generated modes to be sure of %90 in their predictions. However, due to the stochastic nature of generated models, this percentage may not be satisfied all the time.

D. EXPERIMENTAL DESIGN

Forecasting datasets are generally constructed in a time series format, and for algorithms such as ARIMAX and Prophet, forecasting datasets can be used without modification. However, for ML algorithms including GP, the modification of the dataset is a must. For any ML algorithm, it is essential that every input in the dataset has a corresponding output. Modification of the dataset to align with this requirement is necessary. The graphical interpretation of this idea is given in Figure 3.



Time	Observation
1	100
2	110
3	120
4	130
5	140

Time	History (X)	Output (Y)
1	No Info	100
2	100	110
3	110	120
4	120	130
5	130	140

Figure 3. *An example of a time series conversion step*

In Figure 3, a one-step lagged (referred to as day(s) in this paper) supervised version of a hypothetical time series is given. Following this fashion, the Jena dataset was modified so that it contains 10 days of historical data. The same approach was applied again to construct 20 days of historical data. The historical data was utilized to make predictions regarding the weather for 1, 3, and 5 days in advance.

Another crucial aspect of conducting an ML experiment is to divide the dataset into train, test, and validation subsets. Therefore, the overall dataset was divided as %80 training set, %10 test set and %10 validation set. In this study, the validation set were used to generate prediction confidence interval and coverage, as mentioned in Section II. The division scheme is presented in Figure 4.

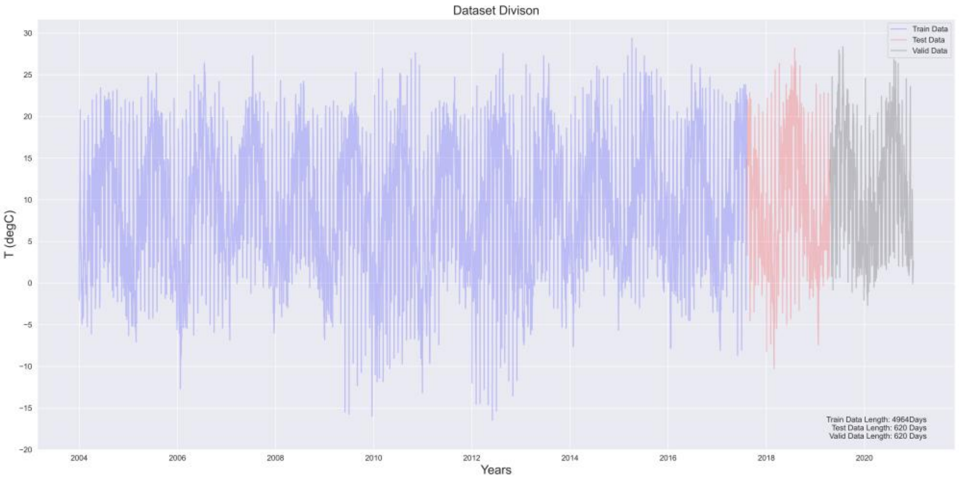


Figure 4. *Dataset division*

4964 days of the dataset were reserved for training, which corresponds to 13.6 years of data. The remaining data was divided equally into test and validation sets, with each set consisting of the same number of days. One of the main distinctions of the experiment is that the models that were generated using the train set were tested and validated monthly. The test set and validation set were divided into 12 months and the models were evaluated for each month separately. Each month has a different characteristic as can be seen in Figure 1. Therefore, conducting monthly evaluations of the models could yield a more comprehensive assessment of their performance.

For GP part, the experiments were carried out for 100 generations, each having 5000 individuals. In each generation, 20 individuals with the highest performance were selected for the next generation. The crossover rate and the mutation rate were %80 and %0.05 respectively. The general settings of the experiments for GP are given in Table 1 and the general schema of the experiments is shown in Figure 5.

Table 1. *Parameters for GP*

Parameters	Value(s)
Non-Terminals [25]	+,*,/,sqrt,max,log,abs,inv,max,min
Terminals	Features in [23]
Generations	100
Population Size	5000
# of Individuals for Next Generation	20
Crossover Probability- Mutation Probability	%80- %0.05

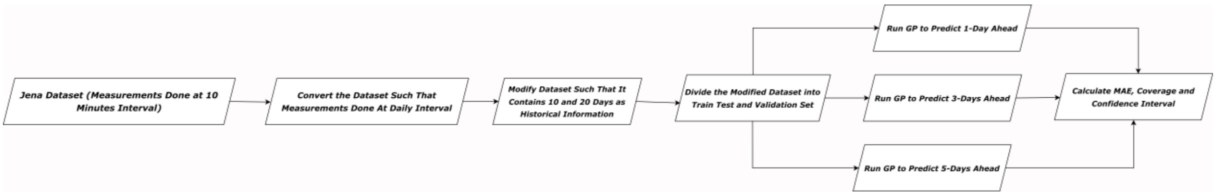

Jena Dataset (Measurements Done at 10 Minutes Interval)
Convert the Dataset Such That Measurements Done At Daily Interval
Modify Dataset Such That It Contains 10 and 20 Days as Historical Information
Divide the Modified Dataset into Train Test and Validation Set
Run GP to Predict 1-Day Ahead
Run GP to Predict 3-Days Ahead
Run GP to Predict 5-Days Ahead
Calculate MAE, Coverage and Confidence Interval

Figure 5. *Overall experiment schema*

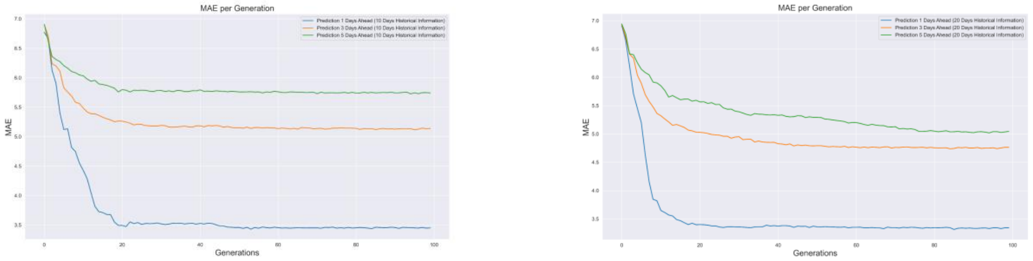
Finally, Mean Absolute Error (MAE) was utilized as a loss function to indicate the error rate of the generated models. The MAE function is given in (1).

$$MAE = \frac{1}{n}\sum_{i=0}^{n} |y_i - \hat{y}_i| \tag{1}$$

where n is the total number of data points, y_i is true continuous value and $\hat{y}_i$ is the forecasted value by the model. All data were normalized before the experiments. Python (version 3.8) programming language was used for the experiments. The next section presents the experimental results and discusses their possible interpretations.

III. RESULTS

The results of the Mean Absolute Error (MAE) computed using 10 and 20 days of historical information for different horizons, are presented in a side-by-side manner in Figure 6 for ease of comparison. It shows MAE results for 100 generations when different amount of past information was used. It also depicts the behavior of the generated models as the forecast horizon advances.



(a) ***(b)***

Figure 6. *Train Performance on GP Programs for Different Historical Information and Forecast Horizons;* ***(a)*** *MAE Results for different forecast horizons using 10 days as information in Training;* ***(b)*** *MAE Results for different forecast horizons using 20 days as information in Training (Best Viewed Online)*

The results showed that the model created by GP to forecast 1 day in advance had a lower error rate compared to the models developed by GP to forecast 3 and 5 days in advance. This statement is valid for both 10 and 20 days of data when they were used as historical information. The MAE was 3.45 when 10 days of historical information were used to forecast 1 day in advance. However, this error rate seemed to increase as the forecast horizon progressed. MAEs for forecasting 3 and 5 days in advance were 5.13 and 5.74, respectively when 10 days of historical information were used.

An important insight that can be derived from the experiment is that there was a substantial discrepancy between forecasting 1 day in advance and 3 days in advance. However, this discrepancy did not arise between forecasting 3 days in advance and 5 days in advance. It can be interpreted that there was not much difference in forecasting between forecasting 3 days and 5 days in advance. The same things could be said for 20 days of information when it was used as past information. MAEs for forecasting 1 day, 3 days and 5 days in advance were 3.43, 4.76, 5.04 respectively. These results suggested that there were no significant changes in MAEs when the amount of information changed. These mentioned results are given in Table 2.

Table 2. *Training Performance of Models Generated by GP (MAE Results)*

Historical Information Forecast Horizon	***10 Days of Information***	***20 Days of Information***
1 Day Ahead	3.45	3.43
3 Days Ahead	5.13	4.76
5 Days Ahead	5.74	5.04

As mentioned before, the models generated by GP were tested monthly. The test and validation set were divided monthly, and the models were tested on each month separately. The test set performance of the models generated by GP is given in Table 3. The generated models are given in Table 4.

Table 3. *Test Set Performance of Models Generated by GP*

	1 Day Ahead Forecast		**3 Days Ahead Forecast**		**5 Days Ahead Forecast**	
	MAE		**MAE**		**MAE**	
	10 Days of Information	**20 Days of Information**	**10 Days of Information**	**20 Days of Information**	**10 Days of Information**	**20 Days of Information**
January	*3.28*	*3.24*	*5.96*	*5.15*	*7.25*	*5.13*
February	*2.93*	*3.47*	*6.34*	*5.31*	*7.79*	*5.70*
March	*3.51*	*3.44*	*5.84*	*5.21*	*6.35*	*4.89*
April	*3.51*	*3.18*	*4.91*	*5.16*	*5.30*	*6.52*
May	*3.95*	*3.59*	*5.77*	*5.59*	*6.42*	*6.13*
June	*4.13*	*3.62*	*4.93*	*4.38*	*5.00*	*3.71*
July	*3.85*	*3.86*	*6.70*	*5.72*	*7.32*	*5.42*
August	*3.25*	*3.18*	*5.51*	*4.91*	*5.53*	*5.14*
September	*3.82*	*3.44*	*4.50*	*4.40*	*4.41*	*4.43*
October	*3.80*	*3.12*	*3.61*	*4.03*	*4.12*	*4.49*
November	*2.82*	*2.78*	*4.70*	*4.03*	*6.17*	*4.60*
December	*2.59*	*2.66*	*5.32*	*4.32*	*6.08*	*4.10*

Although the training set MAE results of the methods gave the general performance, when the month-based approach was used for the test, it was observed that the performance of the models was inconsistent and varied. Table 3 reveals that more information from the past did not necessarily improve the forecasting results. The experiments indicated that forecasting 1 day in advance using 10 and 20 days as information resulted in lower MAE compared to other forecast horizons.

Table 4. *Generated Models by the GP*

	1 Day Horizon (X_n indicates an attribute in the dataset)
10 Days as Information	add(add(X_{182}, X_{183}), add(sqrt(max(log(X_{187}), div(X_{198}, X_{51}))), add(add(X_{182}, add(X_{182}, X_{183})), sqrt(inv(X_{191})))))
20 Days as	add(add(add(sub(add(add(inv(sqrt(X_{251})), X_{381}), X_{383}), log(X_{191})), add(X_{381}, X_{397})), add(X_{381},

Information	X_{397})), X_{381})
	3 Days Horizon (X_n indicates an attribute in the dataset)
10 Days as Information	add(sub(sqrt(inv(X_{71})), log(min(X_{171}, X_{181}))), add(X_{41}, add(add(X_{198}, abs(X_{105})), add(max(X_{118}, X_{197}), X_{62}))))
20 Days as Information	add(add(add(add(add(inv(0.273), inv(sqrt(X_{251}))), add(X_{381}, min(X_{241}, X_{37}))), add(max(X_{297}, X_{397}), add(min(X_{289}, X_{18}), X_{397}))), add(X_{381}, min(X_{261}, X_{18}))), mul(X_{281}, X_{82}))
	5 Days Horizon (X_n indicates an attribute in the dataset)
10 Days as Information	add(X_{197}, sub(add(X_{81}, add(X_2, sqrt(inv(X_{131})))), sub(sub(log(X_{31}), X_{125}), X_{101})))
20 Days as Information	add(add(add(add(add(X_{56}, add(sqrt(inv(0.273)), inv(0.273))), X_{283}), add(add(add(add(inv(0.273), mul(X_{37}, X_{302})), mul(max(X_{25}, X_{117}), X_{345})), mul(X_{61}, X_{261})), add(X_{397}, X_{77}))), X_{18}), min(X_{138}, X_{302}))

However, calculating MAE alone for the test set does not give the overall performance of the methods generated by GP. Confidence and coverage area, that are lower and upper error limits of the models should be indicated. For that purpose, the test set and validation set were used together to calculate the confidence level and coverage area of the models. Table 5 shows these metrics and presents a more vivid picture on the performance of the models. In other words, it shows the overall performance of the models. The lowest bound (L) and the upper bound (U) show the lowest and highest values that the generated models may fluctuate.

Finally, coverage (C) gives the probability of the forecast that will result in between L and U. For example, the prediction results for November differed greatly as the forecasting horizon advances. As the forecasting horizon for November increased, the difference between the lower and upper bounds grew, causing the predictions to resemble random guesses. Similar deductions could be made for all months and forecasting horizons. Although coverage rates were high, the gap between L and U was also very high. Only models for forecasting 1 day in advance produced meaningful predictions.

Table 5. *Test and Validation Set Performance with Error Limits and Coverage*

	1 Day Ahead Forecast		**3 Days Ahead Forecast**		**5 Days Ahead Forecast**	
	MAE with Lower (L) and Upper (U) Limit and Coverage (C)					
	10 Days of Information	**20 Days of Information**	**10 Days of Information**	**20 Days of Information**	**10 Days of Information**	**20 Days of Information**
January	*L=-1.49* *MAE =3.28* *U=13.55* *C = %87*	*L=0.30* *MAE =3.24* *U=13.74* *C = %87*	*L=-4.47* *MAE =5.96* *U=18.86* *C = %93*	*L=-6.80* *MAE =5.15* *U=17.74* *C = %93*	*L=-6.95* *MAE =7.25* *U=22.82* *C = %93*	*L=-8.18* *MAE =5.13* *U=18.15* *C = %93*
February	*L=-1.65* *MAE =2.93* *U=14.55* *C = %86*	*L=-0.64* *MAE =3.47* *U=15.77* *C = %89*	*L=-5.76* *MAE =6.34* *U=19.88* *C = %93*	*L=-5.14* *MAE =5.31* *U=17.20* *C = %93*	*L=-8.10* *MAE =7.79* *U=23.25* *C = %96*	*L=-7.60* *MAE =5.70* *U=19.53* *C = %93*
March	*L=-1.43* *MAE =3.51* *U=14.73* *C = %96*	*L=-0.59* *MAE =3.44* *U=15.77* *C = %93*	*L=-4.19* *MAE =5.84* *U=20.58* *C = %96*	*L=-3.45* *MAE =5.21* *U=17.57* *C = %93*	*L=-7.76* *MAE =6.35* *U=24.20* *C = %96*	*L=-4.79* *MAE =4.89* *U=17.83* *C = %93*
April	*L=1.81* *MAE =3.51* *U=19.22* *C = %100*	*L=2.67* *MAE =3.18* *U=19.02* *C = %100*	*L=-1.89* *MAE =4.91* *U=23.37* *C = %97*	*L=-2.52* *MAE =5.16* *U=23.52* *C = %100*	*L=-3.28* *MAE =5.30* *U=23.68* *C = %95*	*L=-3.55* *MAE =6.52* *U=24.30* *C = %100*
May	*L=3.25* *MAE =3.95* *U=21.12* *C = %96*	*L=3.62* *MAE =3.59* *U=21.43* *C = %95*	*L=2.62* *MAE =5.77* *U=21.57* *C = %83*	*L=1.13* *MAE =5.59* *U=22.70* *C = %90*	*L=-1.35* *MAE =6.42* *U=24.54* *C = %98*	*L=-0.04* *MAE =6.13* *U=23.70* *C = %96*
June	*L=6.25* *MAE =4.13* *U=25.86* *C = %93*	*L=7.30* *MAE =3.62* *U=24.81* *C = %91*	*L=3.17* *MAE =4.93* *U=25.14* *C = %85*	*L=2.17* *MAE =4.38* *U=27.71* *C = %93*	*L=2.32* *MAE =5.00* *U=25.33* *C = %88*	*L=5.55* *MAE =3.71* *U=24.22* *C = %81*
July	*L=6.81* *MAE =3.85* *U=24.54* *C = %88*	*L=7.55* *MAE =3.86* *U=24.39* *C = %88*	*L=1.03* *MAE =6.70* *U=28.54* *C = %93*	*L=2.34* *MAE =5.72* *U=28.35* *C = %98*	*L=0.55* *MAE =7.32* *U=28.62* *C = %93*	*L=3.41* *MAE =5.42* *U=28.34* *C = %95*
August	*L=8.34* *MAE =3.25* *U=21.07* *C = %82*	*L=8.93* *MAE =3.18* *U=21.22* *C = %85*	*L=0.34* *MAE =5.51* *U=27.91* *C = %95*	*L=-0.17* *MAE =4.91* *U=28.95* *C = %100*	*L=3.02* *MAE =5.53* *U=25.61* *C = %93*	*L=3.37* *MAE =5.14* *U=27.02* *C = %98*

September	*L=1.73* *MAE =3.82* *U=21.80* *C = %82*	*L=4.42* *MAE =3.44* *U=21.25* *C = %93*	*L=3.15* *MAE =4.50* *U=21.70* *C = %88*	*L=4.35* *MAE =4.40* *U=20.83* *C = %86*	*L=2.26* *MAE =4.41* *U=23.39* *C = %95*	*L=1.88* *MAE =4.43* *U=23.01* *C = %95*
October	*L=0.82* *MAE =3.80* *U=17.55* *C = %95*	*L=3.77* *MAE =3.12* *U=16.70* *C = %93*	*L=1.22* *MAE =3.61* *U=18.05* *C = %90*	*L=-0.45* *MAE =4.03* *U=18.68* *C = %95*	*L=1.15* *MAE =4.12* *U=19.29* *C = %88*	*L=0.59* *MAE =4.49* *U=18.12* *C = %80*
November	*L=0.26* *MAE =2.82* *U=12.10* *C = %83*	*L=0.04* *MAE =2.78* *U=13.78* *C = %90*	*L=-2.79* *MAE =4.70* *U=16.70* *C = %93*	*L=-3.65* *MAE =4.03* *U=15.51* *C = %90*	*L=-3.26* *MAE =6.17* *U=19.91* *C = %96*	*L=-4.04* *MAE =4.60* *U=16.46* *C = %91*
December	*L=-0.98* *MAE =2.59* *U=13.26* *C = %95*	*L=0.58* *MAE =2.66* *U=13.20* *C = %85*	*L=-5.07* *MAE =5.32* *U=18.42* *C = %96*	*L=-4.76* *MAE =4.32* *U=16.39* *C = %93*	*L=-5.45* *MAE =6.08* *U=21.34* *C = %96*	*L=-5.50* *MAE =4.10* *U=17.26* *C = %89*

One important argument which can be deduced from the experiments is that, for numerical weather predictions, increasing the volume of past information does not affect the results positively. Another argument that can be exposed from the experiments is that GP has the potential for usage in this area, although modification of built-in functions or adding new functions may be necessary. Moreover, it has been shown that the performance of the models can be seen more vividly when ICP is used.

IV. DISCUSSION AND CONCLUSION

Forecasting is a challenging problem in the area of AI, and it is an ongoing research subject for this field. It is mainly based on past observations of the attributes in the dataset, which can be called independent variables, and it is assumed that they represent enough information to predict the dependent variable. Using these assumptions, DL and ML methods which are a subset of AI are utilized. One of the important fields of forecasting is numerical weather prediction. Weather systems are complex systems and difficult to forecast numerically. But the power of ML algorithms may be utilized in this purpose to help researchers at least. However, one important issue for ML algorithms is that they are black box algorithms that are not easily interpretable by humans.

In addition to the interpretability issue of ML in forecasting, forecasting alone is not enough since it is fundamentally prone to errors. For that reason, the confidence and coverage rate of the predictions should be provided with the forecasting models.

In this study, the GP was proposed for interpretability and ICP was proposed for a confidence interval. They were combined to forecast weather temperatures at various horizons using different amount of historical data. Unlike the classical approaches, the testing and validation phases of the experiments were conducted monthly using the Jena dataset. First, the dataset was modified and made suitable for GP, since it works in the format of input-output sequences. Effects of different amount of information, namely 10 and 20 days, were investigated at different horizons, namely 1 day, 3 days and 5 days.

The experimental results showed that even forecasting 1 day in advance was prone to error and increasing the volume of data did not decrease the error. Usage of the ICP approach revealed that the gap between the lower and upper limits of the predictions became so large that the predictions became almost random when forecasting 3 days and 5 days in advance.

We can list some of the possible ideas that could enhance the work presented in this paper. First, the optimal number of past days that maximize information could be investigated. Additionally, the attributes that have the greatest impact on reflecting weather conditions could be examined. This work may be enhanced in the ways mentioned above, which could serve as future work.

V. REFERENCES

[1] Y. Ning, H. Kazemi, and P. Tahmasebi, “A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet,” *Computers & Geosciences*, vol. 164, p. 105126, 2022.

[2] S. J. Taylor and B. Letham, “Forecasting at Scale,” *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[3] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, “Short-term residential load forecasting based on LSTM recurrent neural network,” *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2017.

[4] X. Xiao, H. Mo, Y. Zhang, and G. Shan, “Meta-ANN–A dynamic artificial neural network refined by meta-learning for Short-Term Load Forecasting,” *Energy*, vol. 246, p. 123418, 2022.

[5] S. R. Ghaffari-Razin, A. R. Moradi, and N. Hooshangi, “Modeling and forecasting of ionosphere TEC using least squares SVM in central Europe,” *Advances in Space Research*, vol. 70, no. 7, pp. 2035–2046, 2022.

[6] F. Li and G. Jin, “Research on power energy load forecasting method based on KNN,” *International Journal of Ambient Energy*, vol. 43, no. 1, pp. 946–951, 2022.

[7] M. Lehna, F. Scheller, and H. Herwartz, “Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account,” *Energy Economics*, vol. 106, p. 105742, 2022.

[8] A. Brusaferri, M. Matteucci, P. Portolani, and A. Vitali, “Bayesian deep learning based method for probabilistic forecast of day-ahead electricity prices,” *Applied Energy*, vol. 250, pp. 1158–1175, 2019.

[9] X. Gong, W. Zhang, W. Xu, and Z. Li, “Uncertainty index and stock volatility prediction: evidence from international markets,” *Financial Innovation*, vol. 8, no. 1, pp. 1–44, 2022.

[10] F. Baart, M. van Ormondt, J. van T. de Vries, and M. van Koningsveld, “Morphological impact of a storm can be predicted three days ahead,” *Computers & Geosciences*, vol. 90, pp. 17–23, 2016.

[11] R. Buizza, “Chaos and weather prediction January 2000,” European Centre for Medium-Range Weather Meteorological Training Course Lecture Series ECMWF, 2002.

[12] H. R. Biswas, M. M. Hasan, and S. K. Bala, “Chaos theory and its applications in our real life,” *Barishal University Journal Part*, vol. 1, no. 5, pp. 123–140, 2018.

[13] I. Gad and D. Hosahalli, “A comparative study of prediction and classification models on NCDC weather data,” *International Journal of Computers and Applications*, vol. 44, no. 5, pp. 414–425, 2022.

[14] H. Yang, J. Yan, Y. Liu, and Z. Song, “Statistical downscaling of numerical weather prediction based on convolutional neural networks,” *Global Energy Interconnection*, vol. 5, no. 2, pp. 217–225, 2022.

[15] J. Frnda, M. Durica, J. Rozhon, M. Vojtekova, J. Nedoma, and R. Martinek, "ECMWF short-term prediction accuracy improvement by deep learning," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

[16] G. Xu, K. Lin, X. Li, and Y. Ye, "SAF-Net: A spatio-temporal deep learning method for typhoon intensity prediction," *Pattern Recognition Letters*, vol. 155, pp. 121–127, 2022.

[17] E. K. Burke, G. Kendall (Eds.), "Search methodologies: introductory tutorials in optimization and decision support techniques". *Springer*, 2014.

[18] O. Claveria, E. Monte, and S. Torra, "A Genetic Programming Approach for Economic Forecasting with Survey Expectations," *Applied Sciences*, vol. 12, no. 13, p. 6661, 2022.

[19] E. Christodoulaki, M. Kampouridis, and P. Kanellopoulos, "Technical and sentiment analysis in financial forecasting with genetic programming,", *2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, 2022, pp. 1–8.

[20] S. Cramer, M. Kampouridis, and A. A. Freitas, "Decomposition genetic programming: An extensive evaluation on rainfall prediction in the context of weather derivatives," *Applied Soft Computing*, vol. 70, pp. 208–224, 2018.

[21] M. Sadat-Noori, W. Glamore, and D. Khojasteh, "Groundwater level prediction using genetic programming: the importance of precipitation data and weather station location on model accuracy," *Environmental Earth Sciences*, vol. 79, no. 1, pp. 1–10, 2020.

[22] J. B. Elsner and A. A. Tsonis, "Nonlinear prediction, chaos, and noise," *Bulletin of the American Meteorological Society*, vol. 73, no. 1, pp. 49–60, 1992.

[23] B.Mnassri. (2021, April 06) "*Weather Station Beutenberg Dataset*." [Online]. Available: https://www.kaggle.com/datasets/mnassrib/jena-weather-dataset.

[24] H. Papadopoulos, "Inductive Conformal Prediction: Theory and Application to Neural Networks", *Tools in Artificial Intelligence*. London, United Kingdom: IntechOpen, 2008.

[25] Trevor Stephens (2016) "*API reference – Symbolic Regressor*" [Online]. Available: https://gplearn.readthedocs.io/en/stable/reference.html#reference