New Statistical Indices for Evaluating Model Forecasting Performance

Ioannis Kyriakidis [1, 2], Jaakko Kukkonen[3], Kostas Karatzas[4], George Papadourakis[2] and Andrew Ware[1]

[1] School of Computing and Mathematics,
University of South Wales, Treforest, Pontypridd, CF37 1DL
{kyriakidis@teicrete.gr, andrew.ware@southwales.ac.uk}

[2] Department of Informatics Engineering,
T.E.I. of Crete, Estavromenos, GR-71004 Heraklion, Crete, Greece
{papadour@cs.teicrete.gr}

[3] Finnish Meteorological Institute,
P.O. Box 503, FI 00101, Helsinki, Finland
{jaakko.kukkonen@fmi.fi}

[4] Department of Mechanical Engineering,
Aristotle University, GR-54124 Thessaloniki, Greece
{kkara@eng.auth.gr}

Abstract: A large number of statistical measures have been presented in the literature for the statistical analysis of the agreement of the measured and predicted time-series. The goal of this study was to develop new indices that combine the information contained in several existing measures, making it possible to assess more effectively the quality of the forecasting. The capabilities and limitations of 24 measures that have previously been presented in the literature were studied. The upper and lower bounds of the Confidence Interval were used, in order to include forecasting penalties (relative weights). Results show that by using the proposed new forecasting performance indices we can be more confident in the estimation of the forecasting performance than using a single measure. The proposed new indices would be ideal for a forecasting automated system, because no human interaction is needed to combine the information of other measures.

Keywords: *forecasting performance index, environmental modeling, fuzzy interference systems*

1. INTRODUCTION

The use of mathematical models is essential for understanding, simulating and forecasting the behavior of complex environmental phenomena and systems, like in the case of urban air quality. The evaluation of such forecasting results is necessary regardless of the application domain [1] [2]. In all evaluations, forecasts are compared to relevant observations with the aid of various statistical measures, commonly referred to as indices, which depict various aspects of the differences between forecasted and measured values of the parameters of interest [3] [4] [5] [6].

Jolliffe and Stephenson (2002) [7] define forecast quality as a multidimensional concept described by several different scalar attributes such as overall bias, reliability/calibration, uncertainty, sharpness/refinement, accuracy, association, resolution, and discrimination. All of these attributes provide useful information about the performance of a forecasting system. Thus, no single index is sufficient for forecast evaluation, i.e. for judging and comparing forecast quality [7] [8] [9] [10].

In the current study we compile new indices that improve the evaluation of forecasting results. For this purpose we firstly analyze numerous existing forecasting evaluation indices (FEIs), and we then select the suitable ones to be used for the construction of new statistical forecasting indices. The outcome of this effort is tested in evaluating the forecasting performance of a set of air quality models, and the evaluation results are compared to the ones obtained by using existing evaluation indices. In the rest of the paper we firstly present the methods that we employ and the materials that we use in our study, we then present the results and discuss the performance of the new indices versus the standard indices, and we finalize by drawing our conclusions.

2. MATERIALS AND METHODS

This study was conducted in order to develop new indices that combine the characteristics of existing statistical measures to provide confidence in the forecasting performance estimation than using single existing indices. In order to evaluate the performance of the new indices, we employed artificial neural network models for air quality forecasting.

2.1 STATISTICAL MEASURES SELECTION

As we aimed at developing new indices by making use of existing ones, we selected twenty four (24) statistical measures, which have commonly been used to evaluate the performance of models that produce forecasts of continuous (numerical) variables. For this purpose a literature review was undertaken, resulting in Table 2 which summarizes the most frequently used indices and their basic disadvantages. The most common disadvantages (as reported in literature) were related to a) sensitivity to outliers or to large errors, and b) to a possible division by zero that may occur. Taking these findings into account, we selected a number of existing FEIs to be used as the basis for the generation of the new indices that we wanted to compile, as follows: 1) Index of Agreement (d_r), 2) Legates and McCabe's (E_1), 3) Theil's Inequality Coefficient (U_2), and 4) Berry and Mielke's ($\Re$). The aforementioned indices are described in the following equations:

$$Index\ of\ Agreement\ (d_r) = \begin{cases} 1 - \dfrac{\sum_{i=1}^{n}|F_i - A_i|}{c\sum_{i=1}^{n}|A_i - \bar{A}|}, when \\ \sum_{i=1}^{n}|F_i - A_i| \le c\sum_{i=1}^{n}|A_i - \bar{A}| \\ \dfrac{c\sum_{i=1}^{n}|A_i - \bar{A}|}{\sum_{i=1}^{n}|F_i - A_i|} - 1, when \\ \sum_{i=1}^{n}|F_i - A_i| > c\sum_{i=1}^{n}|A_i - \bar{A}| \end{cases} \quad (1)$$

$$\text{Legates and McCabe's } (E_1) = 1 - \frac{\sum_{i=1}^{n}|A_i - F_i|}{\sum_{i=1}^{n}|A_i - \bar{A}|} \quad (2)$$

$$\text{Theil's Inequality Coefficient } (U_2) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}A_i{}^2}} \quad (3)$$

$$\text{Berry and Mielke's } (\Re) = 1 - \frac{\delta}{\mu} \tag{4}$$

Where,

F_i is the i[th] forecasted value

A_i is the i[th] actual (observed) value

$\bar{A}$ is the mean of the actual values

n is the number of values (forecasted or observed)

c = 2, as suggested by [11] because it balances the number of deviations evaluated within the numerator and within the denominator of the fractional part.

$$\delta = n^{-1} \sum_{i=1}^{n} |F_i - A_i| \tag{5}$$

$$\mu = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} |F_j - A_i| \tag{6}$$

2.2 A METHOD FOR COMPILING NEW FORECASTING EVALUATION INDICES: FUZZY LOGIC

In order to take into account the statistical measures identified as the basis for the construction of the new forecasting performance indices, we wanted to employ a method that imitates human reasoning while being able to map "logic" of the various statistical measures (input space) into the characteristics of the new indices (output space). For this reason we chose to use fuzzy logic [12], which provides a way of processing data by allowing partial set membership (the latter defined on the basis of individual statistical measures), for air quality forecasting.

Fuzzy Logic was selected as a proper method to compile the new statistical indices because a) it allows rapid prototyping (thus appropriate as in our case no prior knowledge on the new indices was available), b) the rules applied can be easily modified (thus supporting the study of different approaches in the construction of the new indices, c) it can encompass great complexity, d) it relates input to output in linguistic terms (i.e. terms easily understood by humans) and e) it has already been used in the past to describe air pollution [13] [14] which is the application domain of our forecasting models.

Fuzzy logic is based on fuzzy sets that form the building blocks for fuzzy conditional rules which have the general form "IF X is A THEN Y is B," where A and B are fuzzy sets. Based upon these rules the decision of fuzzy membership is made. The basic FIS can receive either fuzzy inputs or crisp inputs, but the outputs it produces are commonly fuzzy sets [15]. In this study the Fuzzy Logic Toolbox of MATLAB was used in order to build our FISs.

The most common types of FISs are Mamdani and Sugeno type fuzzy models [16] [17] [18] [19]. Mamdani's FIS is the most commonly used fuzzy methodology and was among the first control systems built using fuzzy set theory. It was proposed in 1975 by Mamdani and Assilian [20]. In this type of FISs, the fuzzy sets that result as the consequent of each rule are combined through an aggregation operator and the resulting fuzzy set is defuzzified to yield the output of the system.

The Sugeno or Takagi-Sugeno-Kang FIS was introduced by Sugeno (1985) [21] and is similar to the Mamdani method. The main difference between Mamdani and Sugeno is the way the crisp output is generated from the fuzzy inputs. While Mamdani uses the technique of defuzzification of a fuzzy output, Sugeno uses weighted average to compute the crisp output.

2.3 EVALUATING THE PERFORMANCE OF THE NEW FORECASTING EVALUATION INDICES: CONFIDENCE INTERVALS

In the frame of our approach, the forecasting performance of a forecasting model is represented its FEIs. But these FEIs are not a set of fixed values but rather (population) parameters characterizing the actual population of forecasted values. This is due to the stochastic nature of model inputs and model parameters on the one hand, and on the inherent difference between real and forecasted values, within each model application. For this reason it is important to characterize not only the performance of a model but also its effectiveness. The latter may be defined as the range of values that contain the FEIs, thus introducing the notion of Confidence Intervals (CIs) in our approach.

Confidence Intervals (CIs) provide a range within which the unknown value of a population parameter is likely to fall. CIs are linked to confidence levels, i.e. the probability value for a population parameter to fall into the CI. The most commonly used confidence levels are 90%, 95% and 99% depending on the application of use. While 95% confidence level is arbitrary, it is traditionally used in applied practice [22]. In this study a confidence level of 95% was used to calculate the CIs.

A resampling method like bootstrapping or cross-validation can be used to generate the sample used for computing the CIs. In the current study, cross-validation is used as a resampling method in order to compute CIs for the FEIs, as it is the one most commonly applied [23], and proved to provide better results than other similar techniques like bootstrapping [23] [24]. In order to compute the CIs with 95% confidence level, we set the bounds (lower and upper) as the 2.5% and 97.5% percentiles, as described by [25]. Depending on the values of the upper and lower bounds of a CI applied in the evaluation procedure, a forecasting model can be characterized as:

(a) Having a relatively high effectiveness (i.e. smaller distance between the CI bounds) to detect the variation of a parameter (in our case the value of the FEIs and thus the forecasting performance), or
(b) Having a relatively low effectiveness (i.e. larger distance between the CI bounds). In that case it can be considered as providing with less information (in terms of forecasting ability).

2.4 POPULATION OF FORECASTING MODELS

A population of forecasting models (with varying forecasting performances-FEIs) was used for the evaluation purposes of our study. In all cases one model type (i.e. ANN-based models) was used, with a) different data sets for training and testing, b) different architectures (in number of neurons on the first hidden layer), and c) varying parameters of the model structure. The population of those models consists of a total of eight different models (**Table 1**) developed in [26]. Those models have different number of inputs (depending on the Dataset) and different number of neurons in the hidden layer (in all cases, only one hidden layer was used).

Table 1: The eight different models that were used in the current study, in relation to the models of [26]

ANN Model Number	**Dataset Number**	**Basic Model Number in** [26]	**Number of inputs**	**Hidden Layer Neurons**
1	1	3	1	2
2	1	3	1	3
3	2	4	9	18
4	2	4	9	23
5	2	4	9	27
6	3	5	10	20

7	3	5	10	25
8	3	5	10	30

2.5 CROSS-VALIDATION

Cross validation (CV) is a popular method applied in order to evaluate the predictive performance of a statistical model [27]. In CV, the available dataset is divided into two segments, one is used to teach or train a model and the other is used to validate the performance of the model. There are several types of CV methods like the Holdout method, k-fold CV, and leave-one-out CV [28]. In this study the 10-fold CV is used in order to 1) measure the predictive performance of the models being developed and 2) compute the CIs of the selected measures to be studied. The next figure shows how the data was divided and used by the forecasting models.

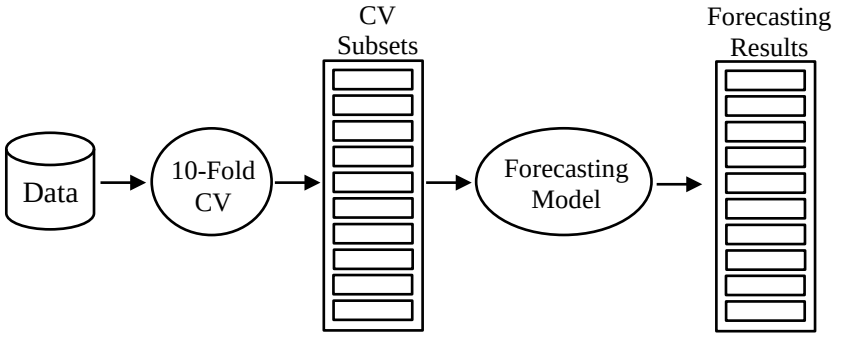


Figure 1: How the data was divided and used by the forecasting models

2.6 PENALIZE FORECASTING PERFORMANCE

The overall forecasting performance (FEIs) is the average of the forecasting performances that were calculated for each CV subset. In order to increase our confidence in the estimation of the forecasting performance relative weights were assigned for each FEI. Those weights aimed at "penalizing" models with relatively low effectiveness, and thus their calculation in based on the bounds of the CIs.

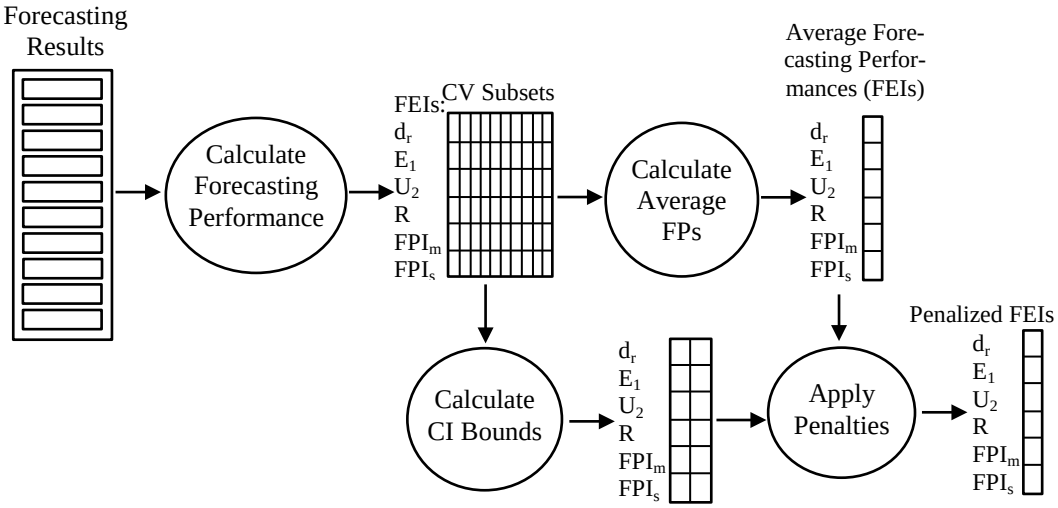


Figure 2: Forecasting Performance Calculation

Table 2: Basic information for the FEIs used in the frame of the current paper

Measure (abbreviation)	**Disadvantage(s)**	
Bias	It provides no measure of the error variance.	[29]
Normalized Bias (NB)	1) Possible division by zero. 2) Positive and negative errors cancel each other out.	
Mean Fractional Bias (MBF)	1) Possible division by zero. 2) Positive and negative errors cancel each other out. 3) The predicted concentration is found in both the numerator and denominator	[30]
Mean Percentage Error (MPE)	1) Possible division by zero. 2) Positive and negative errors cancel each other out.	
Mean Absolute Error (MAE)	Sensitive to outlier errors	
Mean Absolute Percentage Error (MAPE)	Possible division by zero.	
Symmetric Mean Absolute Percentage Error (sMAPE)	Involve division by a number close to zero	[29] [31] [32]
Mean Squared Error (MSE)	Sensitivity to large errors, to large variance of errors and to errors due to outliers	[7]
Normalized Mean Squared Error (NMSE)	Sensitive to extreme values.	
Root Mean Squared Error (RMSE)	Sensitive to large errors, to large variance of errors and on outlier errors	[33]
Linear Correlation Coefficient (r)	Measures only the linear relationships; for instance, a correlation of 0 does not mean zero relationship between two variables.	
Coefficient of Determination (r^2)	1) Based on the linear fit. 2) Sensitive to extreme values.	
Spearman's rank correlation coefficient (r_s)	Simply places the values in numerical order; it pays no regard to the magnitude of the differences between the values.	[34]
Coefficient of Efficiency (E)	Sensitive to extreme values.	[35]
Index of Agreement (d)	Large errors are squared, and thus the influence on the sum-of-squared errors is over-weighted.	[33]
Index of Agreement (d_1)	The overall range of d_1 remained somewhat narrow to resolve adequately the great variety of ways that F can differ from A.	[36]
Index of Agreement (d_r)		[11]
Legates and McCabe's (E_1)		[37]
Legates and McCabe's (E'_1)	Can be used for special cases when season or another time period is available to provide a more appropriate baseline	[37]
Berry and Mielke's ($\mathfrak{R}$)		[38]
Watterson's (M)	The upper and lower bounds it is not well defined.	[39]
Factor of Exceedance (FOEX)	Cannot distinguish an under-prediction than a perfect fit	[40]
Theil's Inequality Coefficient (U_1)	It has a little or no value as a forecasting accuracy index.	[41]
Theil's Inequality Coefficient (U_2)		[42]

We refer those relative weights as "penalties", because they are calculated so as to decrease the forecasting performance of a model, but to not affect (reward) the forecasting models with relatively high effectiveness. The penalties are defined in Eq. 7 where the "Penalty Cancel Level" defines the decrement size of the penalty effect. A small value of Penalty Cancel Level (PCL) will provide a large penalty effect and a large value of the PCL will provide small penalty effect. We applied three different PCLs, equal to 0.5, 1 and 1.5respectivelly.

$$penalty = 1 - (\text{NormalizedDistance}^{PenaltyCancelLevel}) \tag{7}$$

Where,

distance = (CI upper bound – CI lower bound)

$$NormalizedDistance = \begin{cases} 1, when\ distance > 1 \\ distance, when\ distance \leq 1 \end{cases}$$

In Eq. 7 the distance was normalized to a maximum value of 1. The reason we used a value range between 0 and 1 was because, a) some of the selected FEIs converge to infinite values, and b) this value range is more appropriate for comparing performances and drawing conclusions.

Equations 8 and 9 show how the penalty was assigned in the forecasting performance depending on the FEI's nature. Equation 8 is used in (d_r, E_1 and $\Re$) measures, in which a low value indicates low forecasting performance and a high value indicates high forecasting performance. On the other hand, Equation 9 is used in (U_2) measure, in which a low value indicates high forecasting performance and a high value indicates low forecasting performance.

$$PenalizedVariable_1 = \text{var}iable * penalty \tag{8}$$

$$PenalizedVariable_2 = \text{var}iable / penalty \tag{9}$$

3. THE NEW FORECASTING PERFORMANCE INDICES

The performance of any forecasting model (like the ones used for air quality) is commonly evaluated with the aid of several forecasting evaluation measures (statistical indices). On this basis, models can be compared and thus the best model can be selected. As there is no formal procedure to follow, this means that for the same model results, different researchers may select different model(s) as the best in terms of forecasting, depending on (a) the employed statistical indices and (b) the criteria used for interpreting the values of these indices. In an effort to overcome this problem we developed two new forecasting evaluation indices, denoted as FPI_m and FPI_s for Mamdani-type and Sugeno-type FIS respectively. For this reason, we firstly present the way in which the FISs were build and we present the evaluation process that was used to evaluate the Forecasting Performance Indices (FPI).

3.1 BUILDING OF THE FUZZY INFERENCE SYSTEM

3.1.1 SPECIFY THE INPUTS AND OUTPUTS

The four measures selected as inputs (input space) for the FIS are the ones indicated in chapter 2.1, i.e. the Index of Agreement (d_r), the Legates and McCabe's (E_1), the Theil's Inequality Coefficient (U_2) and Berry and Mielke's ($\Re$). The output of the FIS will be the Forecasting Performance scaled in five levels, as suggested by the Altman method). The Altman's Kappa benchmark uses the following five scales: a. Poor (< 0.20), b. Fair (0.21 to 0.40), c. Moderate (0.41 to 0.60), d. Good (0.61 to 0.80), and e. Very Good (0.81 to 1.00). Although this benchmark is developed to be used with the Kappa coefficient, it is often used in practice with other statistical indices as well [43].

3.1.2 DETERMINE THE MEMBERSHIP FUNCTION FOR EACH INPUT AND OUTPUT

As a next step we define the membership functions associated with each of the inputs and variables (scale of the output). Figure 3 shows the membership function for the inputs d_r, E_1 and $\Re$, in which the Gaussian curve membership function was used with the five Altman's Kappa levels ranges from 0 (as Poor) to 1 (Very Good). In addition, Figure 4 shows the membership function for the input U_2 in which the same function was used but with level ranges from 0 (as Very Good) to 1 (Poor). The selected statistical indices can receive values outside the range [0, 1], that were mapped to the minimum (zero) and the maximum (one) values in an appropriate way. Figure 5 and Figure 6 shows the Mamdani and Sugeno FIS that were developed in this study.

3.1.3 FIS RULES

In the last step we constructed the rules of the FIS (presented hereafter). The basic idea behind these rules is to support the mapping process between the input and the output space. Thus, the FIS rules were defined to relate each input (forecasting evaluation index) with the output (forecasting performance). One of the reasons that led us to this solution was that there are no generally accepted rules to relate the selected measures, a) with each other, and b) with the forecasting performance. An additional reason that led us to the aforementioned solution was that it was not possible to create a rule for each combination of input-output. In that case a very large number of rules would be created (a total of $performance\ levels^{number\ of\ indices} = 5^4 = 625$), and also it would be difficult to select the appropriate output for each input. The FIS rules are the result of the 4 indices compared with the 5 levels of performance (a total of 20 rules), as detailed below (x receives the value poor, fair, moderate, good or very good).

- If ($\Re$ is x) then (Forecasting Performance is x) (1)
- If (U_2 is x) then (Forecasting Performance is x) (1)
- If (E_1 is x) then (Forecasting Performance is x) (1)
- If (d_r is x) then (Forecasting Performance is x) (1)

The numbers in the parentheses represent weights that are applier to each rule. Every rule has a weight (a number between 0 and 1), which is applied to the number given by the antecedent. In all our rules the weight has a value of one (has no effect at all on the mapping process).

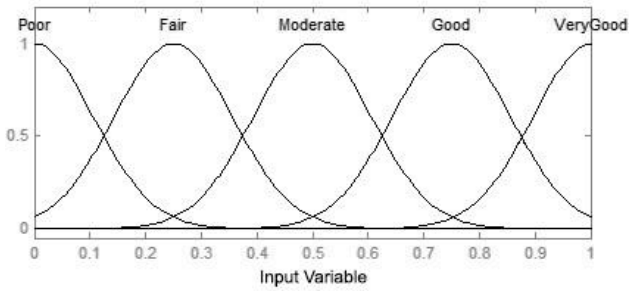


Figure 3: The membership function for the inputs (d_r, E_1 and $\Re$), as also of the output.

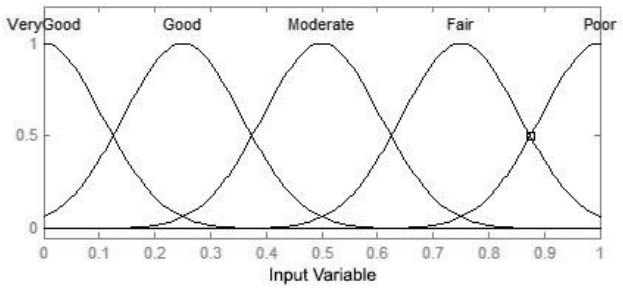


Figure 4: The membership function for the input U_2.

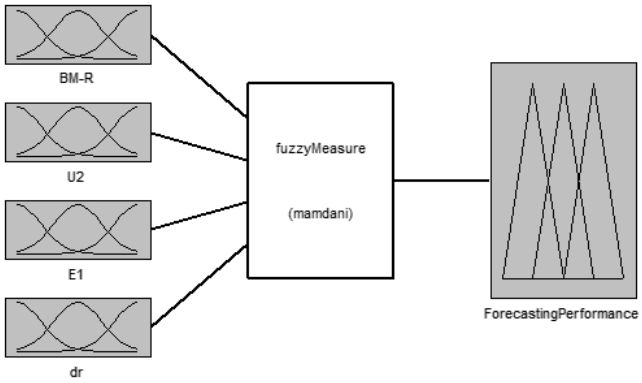


Figure 5: The Mamdani FIS

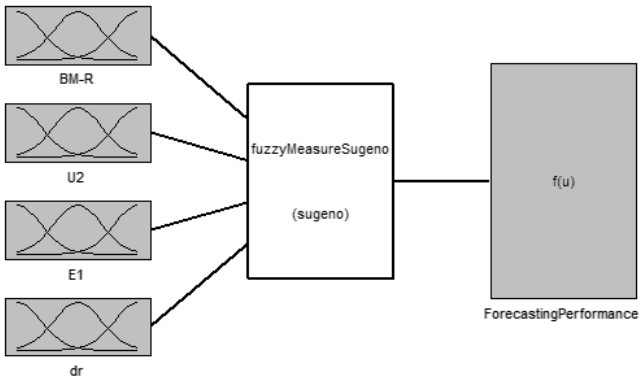


Figure 6: The Sugeno FIS

4. RESULTS AND DISCUSSION

To evaluate the new FEIs (FPI_m and FPI_s), we use them for a population of models with varying forecasting performances. In these models, our aim was to forecast the Common Air Quality Index (CAQI) numerical values (dependent variable). For the sake of our study, we generated a population of model results by initiating and running eight ANN-based forecasting models 1000 times each. We have then computed the output values of the four selected FEIs as well as the output values of the two new indices being introduced, in order to study the consistency of their behaviors. This computation is repeated for each one of the PCL used (0.5, 1 and 1.5) in order to also study its influence in the evaluation of the forecasting performance.

Firstly, the forecasting performance was calculated based on each one of the indices under study. Subsequently, we have calculated the percentage where each one of the eight forecasting models has been identified as a) best forecasting model, or b) best or second best forecasting model. The "best" forecasting model is the one with the highest value for each one of the FEIs. The "second best forecasting model" is the one with the next higher performance for each one of the studied indices.

Figure 7 presents the percentages of the cases, in which each model has been evaluated to be the best, by using different PCLs. These percentages are different, depending on which FEI (d_r, E_1, U_2 etc.) was used for selecting the best model. **Figure 8** presents the mean values of each index in the cases where each forecasting model was identified as best, when no PCLs were applied. In **Figure 7**.a, **Figure 7**.d and **Figure 8**, the models 6-8 have zero values because they were not identified as best in any of the cases. From **Figure 7**.a and **Figure 8** it is clear that d_r and E_1 FEIs lead to the same models being identified as best, even when their mean performance values are different. This does not come as a surprise as, according to [11], these two measures are related. From **Figure 7**.a it is evident that Model 3 is identified as the best forecasting model on the basis of all FEIs under study.

From **Figure 7** we observe that, if we do not use "penalties", Model 3 is identified as the best model by all FEIs, whereas, when penalties are employed (with PCL of 0.5), the U_2 and d_r measures lead to a different model being identified as best. This suggests that CI's distance is not only influenced by the forecasting model but also by the FEI that was selected. Thus, it is crucial to select a reliable forecasting performance measure. It should be mentioned that measures E_1, $\Re$ and our new FPIs (FPI_m and FPI_s) identified Model 3 as the best model in all cases (with different PCLs). This demonstrates that those measures are more stable (in terms of consistency in the results when using penalties) in comparison with measures U_2 and d_r.

It is clear that when penalties are employed, the high performance of some models deteriorates, because of their CI's in comparison to the other models. This suggests that we cannot be confident for the evaluation of the forecasting performance of a model, even when it is accompanied by high FEIs. Thus, we can identify the necessity to use a CI in order to penalize the forecasting performance measures and increase our confidence in the estimation of the forecasting performance. By observing **Figure 7**.a and **Figure 7**.b, we can indicate whether a FEI is stable, if the percentages of the cases, in which each model has been evaluated to be the best (by not using a PCL), are the same as in the case where a PCL is used. In this manner we can use the following equation to calculate the stable percentage of a FEI.

$$Stable\ Percentage = \frac{\sum_{i=1}^{N}(100 - |m_i(0) - m_i(0.5)|)}{N} \quad (10)$$

Where,

$m_i(x)$ is the percentage of the cases in which each model has been evaluated to be the best, by using a FEI (m) for model i, with PCL equal to x.

N is the total number of models (in our case N = 8)

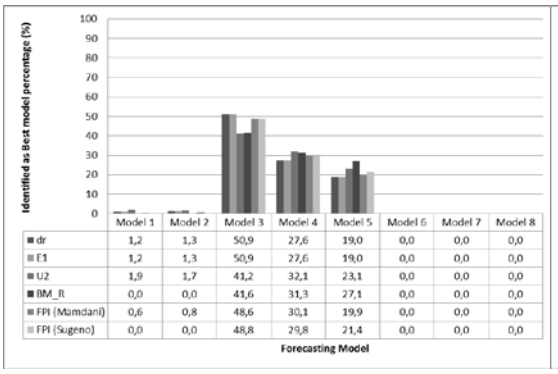
It is useful to identify also the second best model, so we can recommend an alternative solution and in general not be strict with the best model. For that purpose, **Figure 9** was created in order to present the percentages of the cases, in which each model has been evaluated to be the best or the second best, by using different PCLs. The stable percentages of each FEI are: $\Re$=98%, E_1=93.7%, FPI_s=91%, FPI_m=90%, d_r=87.7% and U_2=80.6%. From **Figure 9** and from the aforementioned stable percentages we can see that measure $\Re$ is the most stable FEI (98%) in terms of varying penalty in comparison to the other FEIs, because its results are not affected by the increase of the penalty in comparison with the other indices.

The new FPIs (both FPI_m and FPI_s) were designed in such a way that they take into account all selected measures in a balanced way (by using the same weight value in the fuzzy rules). These indices are better compared to single measures, in respect of confidence in the estimation of the forecasting performance, because they:

1. Use a combination of measures (while we have shown that we cannot be confident in any single measure's estimation)
2. Use the CI in order to penalize the forecasting performance (and thus increase the confidence in the obtained results)

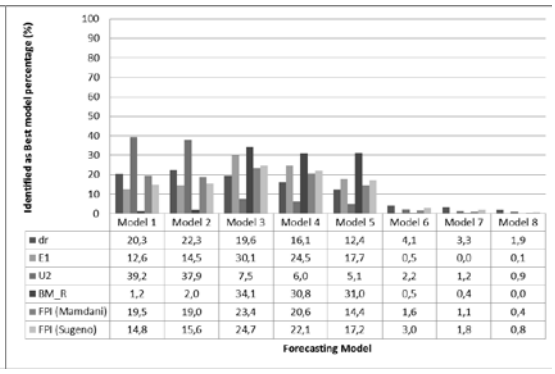
3. Are stable to an acceptable level 90% - 91%, which is better than: d_r (87.7%) and U_2 (80.6), but worse than: $\Re$(98%) and E_1 (93.7).
4. Can potentially make the evaluation process of forecasting models more straightforward and robust
5. Can be used in a forecasting system, for automatically selecting and switching to a different operational forecasting model.

By comparing the results obtained for the two new FPIs, Mamdani-type (FPI_m) and Sugeno-type (FPI_s), it is evident that both demonstrate similar behavior, with Sugeno-type FIS being slightly more stable (by 1%). This is because, Sugeno-type FIS is not affected by the increase of the penalty in comparison with the Mamdani-type FIS.



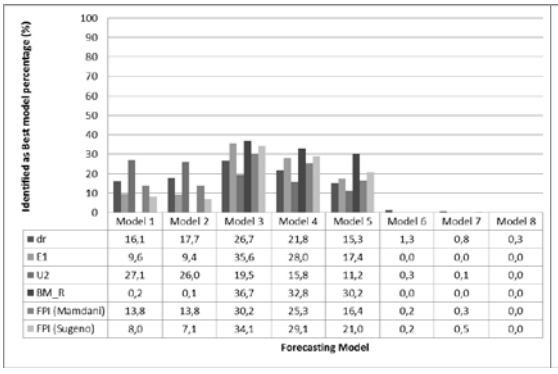
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
dr	1,2	1,3	50,9	27,6	19,0	0,0	0,0	0,0
E1	1,2	1,3	50,9	27,6	19,0	0,0	0,0	0,0
U2	1,9	1,7	41,2	32,1	23,1	0,0	0,0	0,0
BM_R	0,0	0,0	41,6	31,3	27,1	0,0	0,0	0,0
FPI (Mamdani)	0,6	0,8	48,6	30,1	19,9	0,0	0,0	0,0
FPI (Sugeno)	0,0	0,0	48,8	29,8	21,4	0,0	0,0	0,0

Figure 7.a: No penalty cancel level was used



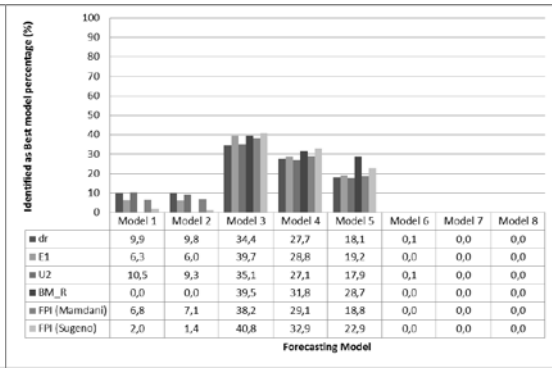
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
dr	20,3	22,3	19,6	16,1	12,4	4,1	3,3	1,9
E1	12,6	14,5	30,1	24,5	17,7	0,5	0,0	0,1
U2	39,2	37,9	7,5	6,0	5,1	2,2	1,2	0,9
BM_R	1,2	2,0	34,1	30,8	31,0	0,5	0,4	0,0
FPI (Mamdani)	19,5	19,0	23,4	20,6	14,4	1,6	1,1	0,4
FPI (Sugeno)	14,8	15,6	24,7	22,1	17,2	3,0	1,8	0,8

Figure 7.b: A penalty cancel level of 0.5 was used



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
dr	16,1	17,7	26,7	21,8	15,3	1,3	0,8	0,3
E1	9,6	9,4	35,6	28,0	17,4	0,0	0,0	0,0
U2	27,1	26,0	19,5	15,8	11,2	0,3	0,1	0,0
BM_R	0,2	0,1	36,7	32,8	30,2	0,0	0,0	0,0
FPI (Mamdani)	13,8	13,8	30,2	25,3	16,4	0,2	0,3	0,0
FPI (Sugeno)	8,0	7,1	34,1	29,1	21,0	0,2	0,5	0,0

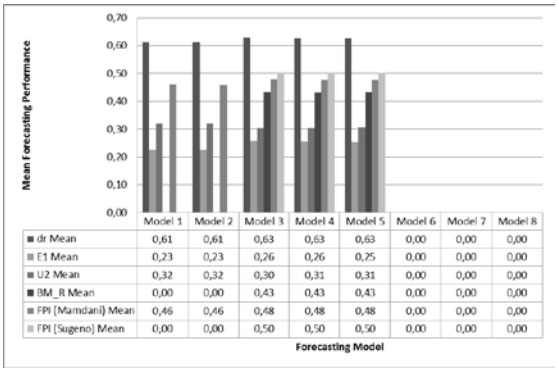
Figure 7.c: A penalty cancel level of 1 was used



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
dr	9,9	9,8	34,4	27,7	18,1	0,1	0,0	0,0
E1	6,3	6,0	39,7	28,8	19,2	0,0	0,0	0,0
U2	10,5	9,3	35,1	27,1	17,9	0,1	0,0	0,0
BM_R	0,0	0,0	39,5	31,8	28,7	0,0	0,0	0,0
FPI (Mamdani)	6,8	7,1	38,2	29,1	18,8	0,0	0,0	0,0
FPI (Sugeno)	2,0	1,4	40,8	32,9	22,9	0,0	0,0	0,0

Figure 7.d: A penalty cancel level of 1.5 was used

Figure 7: The percentages of the cases, in which each model has been evaluated to be the best, by using different penalty cancel levels.



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
dr Mean	0,61	0,61	0,63	0,63	0,63	0,00	0,00	0,00
E1 Mean	0,23	0,23	0,26	0,26	0,25	0,00	0,00	0,00
U2 Mean	0,32	0,32	0,30	0,31	0,31	0,00	0,00	0,00
BM_R Mean	0,00	0,00	0,43	0,43	0,43	0,00	0,00	0,00
FPI (Mamdani) Mean	0,46	0,46	0,48	0,48	0,48	0,00	0,00	0,00
FPI (Sugeno) Mean	0,00	0,00	0,50	0,50	0,50	0,00	0,00	0,00

Figure 8: The mean values of each measure in the cases, for which each forecasting model was identified as best, when no penalty cancel levels were applied.

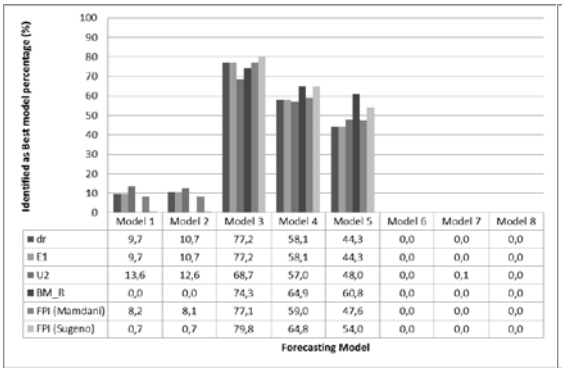


Figure 9.a: No penalty cancel level was used

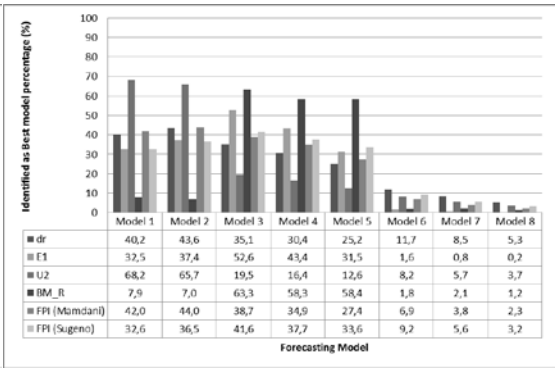


Figure 9.b:penalty cancel level of 0.5

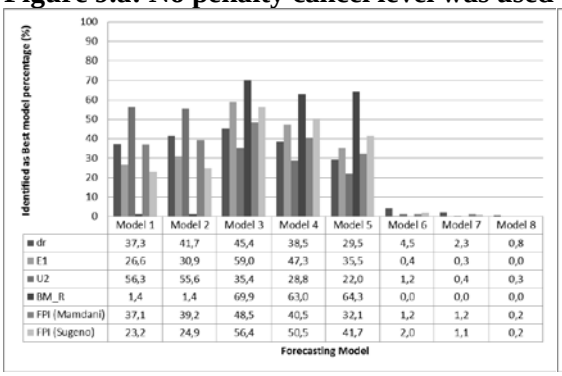


Figure 9.c: penalty cancel level of 1

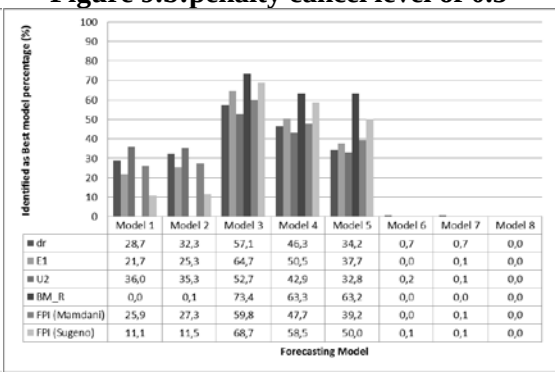


Figure 9.d: penalty cancel level of 1.5

Figure 9: The percentages of the cases, in which each model has been evaluated to be the best or the second best, by using different penalty cancel levels.

5. CONCLUSION

The paper introduces two new forecasting performance indices, which combine the characteristics of several statistical evaluation measures. The relative difference between the construction of the new indices (FPI_m and FPI_s) is the type of fuzzy inference system employed for the construction of each index (Mamdani and Sugeno, respectively). In addition, in order to increase our confidence in the estimation of the forecasting performance of each forecasting model, relative weights (referred to as penalties), based on the bounds of the confidence intervals were assigned on the forecasting performance of each model.

In order to evaluate the new forecasting performance indices, numerical simulations have been performed by using artificial neural network models for air quality forecasting. Results show that we cannot be confident in any single measure's estimation, even when a model is estimated of having a high forecasting performance. Thus, it is important to make use of a confidence interval in order to penalize the forecasting performance measures and thus increase our confidence in the estimation of the forecasting performance.

Results demonstrate that the $\Re$ measure was more stable (98%) in terms of varying the penalty than the other selected measures, because it was not affected by the increase of the penalty in comparison with the other measures. Our new forecasting performance indices (FPI_m and FPI_s) were not the most stable measures. This may be related to the fact that simple FISs (with one antecedent per rule) were used. The increase of the stability of the proposed indices will be investigated in future work.

The new proposed indices a) are stable to an acceptable level (90% - 91%), b) provide a combination of several measures, that increases our confidence in the estimation of the forecasting performance and standardize the interpretation, c) the forecasting performance becomes comparable with the results of other studies (because it is a percentage value), and d) the forecasting performance is scaled in five levels, so that the results can be easier interpret-

ed. Thus, by using the proposed new forecasting performance indices (both FPI_m and FPI_s) in combination with the use of confidence intervals for penalizing the forecasting performance, we can be more confident for the estimation of the forecasting performance of a model than by using any single measure. Consequently, it can be considered that the new forecasting performance indices are appropriate for automated operational forecasting systems.

6. REFERENCES

[1]. A. H. Murphy and R. L. Winkler, "A general framework for forecast verification," *Monthly Weather Review,* vol. 115, no. 7, pp. 1330-1338, 1987.

[2]. C. A. Doswell III, "Verification of forecasts of convection: Uses, Abuses, and Requirements," Avoca Beach, New South Wales, Australia, 1996.

[3]. K. D. West, "Forecast Evaluation," in *Handbook of Economic Forecasting*, E. Graham, G. Clive and A. Timmermann, Eds., Elsevier B.V., 2006, pp. 99-134.

[4]. F. X. Diebold and J. A. Lopez, "Forecast Evaluation and Combination," in *Handbook of Statistics 14: Statistical Methods in Finance,* G. S. Maddala and C. R. Rao, Eds., Amsterdam, North-Holland, Butterworth Heinemann, 1996, pp. 241-268.

[5]. S. Laurent and F. Violante, "Volatility forecasts evaluation and comparison," *WIREs Computational Statistics,* vol. 4, no. 1, pp. 1-12, 2011.

[6]. T. E. Clark and M. W. McCracken, "Advances in Forecast Evaluation," Research Division: Federal Reserve Bank of St. Louis, 2011.

[7]. I. T. Jolliffe and D. B. Stephenson, Forecast Verification: A Practitioner's Guide in Atmospheric Science, Wiley and Sons Ltd., 2002.

[8]. A. P. Morse, F. J. Doblas-Reyes, M. B. Hoshen, R. Hagedorn and T. N. Palmer, "A forecast quality assessment of an end-to-end probabilistic multi-model seasonal forecast system using a malaria model," *Tellus,* vol. 57, no. 3, pp. 464-475, 2005.

[9]. W. M. Briggs and R. A. Levine, "Wavelets and Field Forecast Verification," *American Meteorological Society,* vol. 125, pp. 1329-1341, 1997.

[10]. S. J. Mason and D. B. Stephenson, "How Do We Know Whether Seasonal Climate Forecasts Are Any Good?," in *Seasonal Climate: Forecasting and Managing Risk*, A. Troccoli, M. Harrison, D. Anderson and S. J. Mason, Eds., Springer Netherlands, 2008, pp. 259-289.

[11]. C. J. Willmott, S. M. Robeson and K. Matsuura, "A refined index of model performance," *International Journal of Climatology,* vol. 32, no. 13, p. 2088–2094, 2011.

[12]. L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages," *Computers & Mathematics with Applications: Special Issue on Computational Linguistics,* vol. 9, no. 1, pp. 149-184, 1983.

[13]. K. Karatzas, "A fuzzy logic approach in Urban Air Quality Management and Information Systems," in *4th International Conference on Urban Air Quality Measurement, Modelling and Management,* Prague, Czech Republic, 2003.

[14]. J. Y. Yadav, V. Kharat and A. Deshpande, "Fuzzy Description of Air Quality: A Case Study," Banff, Canada, 2011.

[15]. S. N. Sivanandam, S. Sumathi and S. N. Deepa, Introduction to Fuzzy Logic using MATLAB, Berlin: Springer, 2007.

[16]. S. A. Alshalaa and E. M. Issmail, "Comparison of Mamdani and Sugeno Fuzzy Interference Systems for the Breast Cancer Risk," *World Academy of Science, Engineering and Technology: International Journal of Computer Science and Engineering,* vol. 7, no. 10, pp. 840-844, 2013.

[17]. A. Kaur and A. Kaur, "Comparison of Mamdani-Type and Sugeno-Type Fuzzy

Inference Systems for Air Conditioning System," *International Journal of Soft Computing and Engineering (IJSCE),* pp. 323-325, 2012.

[18]. S. Mansi and B. Gajanan, "Comparison of Mamdani and Sugeno Inference Systems for Dynamic Spectrum Allocation in Cognitive Radio Networks," *Wireless Personal Communications,* vol. 71, no. 2, pp. 805-819, 2013.

[19]. V. Kansal and A. Kaur, "Comparison of Mamdani-type and Sugenotype FIS for Water Flow Rate Control in a Rawmill," *Scientific & Engineering Research,* vol. 4, no. 6, pp. 2580-2584, 2013.

[20]. E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies,* vol. 7, no. 1, pp. 1-13, 1975.

[21]. M. Sugeno, Industrial applications of fuzzy control, 1st ed., New York: Elsevier Science Ltd, 1985.

[22]. D. G. Rees, Foundations of Statistics, Chapman and Hall/CRC, 1987.

[23]. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *14th international joint conference on Artificial intelligence*, Montreal, Quebec, Canada, 1995.

[24]. U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics,* vol. 20, no. 3, p. 374–380, 2004.

[25]. D. T. Kaplan, Resampling Stats in MATLAB, Arlington, Virginia: Resampling Stats, Inc, 1999.

[26]. I. Kyriakidis, K. Karatzas, J. Kukkonen, G. Papadourakis and A. Ware, "Evaluation and analysis of artificial neural networks and decision trees in forecasting of common air quality index in Thessaloniki, Greece," *Engineering Intelligent Systems,* vol. 21, no. 2/3, pp. 111-124, 2013.

[27]. P. Refaeilzadeh, L. Tang and H. Liu, "Cross-Validation," in *Encyclopedia of Database Systems*, L. Liu and T. M. Özsu, Eds., Springer Science+Business Media, LLC, 2009, pp. 532-538.

[28]. S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys,* vol. 4, p. 40–79, 2010.

[29]. S. J. Armstrong, Long-Range Forecasting: From Crystal Ball to Computer, New York: Wiley-Interscience, 1985, pp. 332-361.

[30]. Q. Ying, M. P. Fraser, R. J. Griffin, J. Chen and M. J. Kleemanm, "Verification of a source-oriented externally mixed air quality model during a severe photochemical smog episode," *Atmospheric Environment,* vol. 41, no. 7, pp. 1521-1538, 2007.

[31]. S. Makridakis and M. Hibon, "The M3-Competition: results, conclusions and implications," *International Journal of Forecasting,* vol. 16, no. 4, pp. 451-476, 2000.

[32]. R. R. Andrawis and A. F. Atiya, "A New Bayesian Formulation for Holt's Exponential Smoothing," *Journal of Forecasting,* vol. 28, pp. 218-234, 2009.

[33]. C. J. Willmott, "On the validation of models," *Physical Geography,* vol. 2, no. 2, pp. 184-194, 1981.

[34]. N. Thornes, "Using Spearman's Rank Correlation Coefficient in Coursework," Geofile Online, 2006.

[35]. J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models, Part I, A discussion of principles," *Journal of Hydrology,* vol. 10, no. 3, pp. 282-290, 1970.

[36]. C. J. Willmott, S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell and C. M. Rowe, "Statistics for the Evaluation and Comparison of Models," *Journal of Geophysical Research,* vol. 90, no. C5, pp. 8995-9005, 1985.

[37]. D. R. Legates and G. J. McCabe, "Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation," *Water Resources Research,* vol. 35,

no. 1, pp. 233-241, 1999.

[38]. K. J. Berry and P. W. Mielke, "A Family of Multivariate Measures of Association for Nominal Independent Variables," *Educational and Psychological Measurement,* vol. 52, no. 1, pp. 41-55, 1992.

[39]. I. G. Watterson, "Non-Dimensional Measures of Climate Model Performance," *International Journal of Climatology,* vol. 16, no. 4, pp. 379-391, 1996.

[40]. R. S. Sokhi, R. San Jose, N. Kitwiroon, E. Fragkou, J. L. Perez and D. R. Middleton, "Prediction of ozone levels in London using the MM5-CMAQ modelling system," *Environmental Modelling and Software,* vol. 21, no. 4, pp. 566-576, 2006.

[41]. H. Theil, Economic Forecasts and Policy, North Holland Publishing Company, 1958.

[42]. H. Thiel, Applied Economic Forecasting, Elsevier Science, 1966.

[43]. G. L. Kilem, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters, Advanced Analytics, LLC, 2012.