**2018**
**Mathematical Contest in Modeling (MCM/ICM) Summary Sheet**

# CAFE: Characterization, Analysis, Forecasting, and Evaluation of Energy Profile

**Summary**

As sustainability comes in the spotlight, *renewable energy* receives extensive study. One fresh idea to it in U.S. is the *Interstate Energy Compact*, aiming at promoting collaboration for energy use among states. In this paper, we formulate concrete objectives and actions on the use of renewable energy for such a compact among California (CA), Arizona (AZ), New Mexico (NM) and Texas (TX).

We first build **CAFE**, a novel framework for Characterization, Analysis, Forecast and Evaluation on the energy profile (EP) of a state. We constitute EP with 20 items in a multifaceted manner selected and aggregated from the 605 variables in the provided data. We utilize the **Gaussian Process Regression (GPR)** model to characterize the basic evolving trends, strong fluctuations and random noise level of different EP time series in each state from 1960 to 2009. We combine **Gray Relational Analysis** and **Kendall Rank** to measure the similarity of EPs among states in value and tendency respectively, and use both **Pearson Correlation Coefficient** and **Partial Relational Coefficient** to unveil outer influential factors on the similarity. We consider 7 criteria concerning renewable energy, endow them with different importance by the **Entropy Weight Method**, integrate them into an ultimate score based on **TOPSIS** method, and finally judge CA having the 'best' profile. We pioneeringly formulate an ARMA-GPR Hybrid model for EP prediction of 2025 and 2050, which adaptively suits itself to short-term and long-term prediction. We further propose a sliding window mechanism and a 'Look-Ahead' refinement approach incorporating more information in prediction.

We then determine goals and approaches for renewable energy usage in 2025 and 2050. We discover the predominance of renewable energy consumption (RC) in its usage, and thus model the goal setting problem as a multi-objective optimization over RC. Results show that RC targets for 2025/2050 are respectively CA: 988787/1227217, AZ: 148028/175154, NM: 30120/36359, TX: 283871/371206 (billion Btu). We further propose three actions to stimulate RC in AZ, NM, and TX, and show that CA can export more renewable energy to help the other three states in enhancing RC.

We finally conduct sensitivity analysis, dissect pros and cons of our model and present a memo of our work to the state governors.

**Keywords**: CAFE; Energy Profile; Correlation Analysis; ARMA-GPR Model; Multi-Objective

# Contents

# 1   Introduction

With the mounting heterogeneity of energy infrastructures and demands in different regions, centralized regulation of energy can no longer meet the needs for economic development in every state in U.S. To confront this, two well-known interstate compacts: Western Interstate Energy Compact (1970) and South States Energy Compacts (1978), were proposed to make for state economic enhancement in a decentralized and cooperative manner[1].

In 21st century, as sustainability becomes under great concern, the same issue of decentralization falls upon the renewable energy management. Currently, the state governors of California, Arizona, New Mexico, and Texas are planning to formulate a renewable energy interstate compact. Specifically, they need

- an energy profile[1] for states to characterize and evaluate their renewable energy use, and make predictions for it in 2025 and 2050;

- a list of goals written in the compact for renewable energy usage in 2025 and 2050, and at least three actions to achieve these goals.

A slew of prior arts have shed light on the energy prediction: Cammarano et. al [3] proposed *Pro-Energy*, a prediction model of future available energy for wireless sensor networks, Virote and Neves-Sirva [4] built an energy consumption model based on stochastic Markov models to forecast energy saving. However, above works zoomed in on prediction within days, far less than intervals of years as the compact requires. Moreover, no prior works have ever raised any quantified goal or action for renewable energy in an interstate energy compact.

In this work, we propose a novel framework, named **CAFE** (Characterization, Analysis, Forecasting, and Evaluation of Energy Profile), to fill in the gap above. The framework of CAFE is shown in Figure 1.
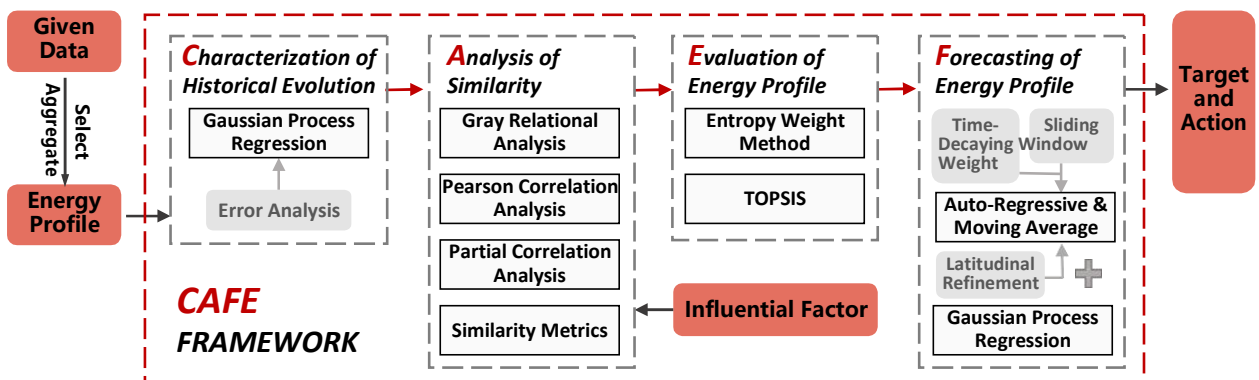


Figure 1: Framework of CAFE (Characterization, Analysis, Forecasting, and Evaluation of Energy Profile)

CAFE framework in Fig. 1 can be summarized as the following steps:

---

[1]The energy profile primarily consists of energy production, consumption, import, and exports [2].

- **Characterization:** We create an energy profile for each state as an integration of 20 typical items, and build *Gaussian Process Regression (GPR)* model to show the evolution of energy profile.

- **Analysis:** utilize Correlation Analysis to identify the similarities of energy profile among states with the potential causes including geography, climate, industry, etc.

- **Evaluation:** We apply *Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)* method to form an integrated criterion based on seven refined criteria.

- **Forecast:** We novelly propose an *ARMA-GPR* model which adaptively suits well in both short-term and long-term energy profile prediction. This model not only integrates *GPR* and *ARMA* by allocating a time-decaying weight, but also incorporates the similarity among states.

- **Targets and Actions:** We utilize multi-objective programming to determine quantified targets for renewable energy consumption in four states in 2025 and 2050, and provide actions stimulating the renewable energy use in AZ, NM, and TX based on our prediction.

# 2 Assumptions

First and foremost, we make some basic assumptions and explain their rationales.

**Assumption 1.** *Each state attaches great importance to renewable energy usage and sets a shared goal to enhance the management and development of renewable energy.*

This assumption is the premise of our work since only when each state endeavors to increase the usage of renewable energy, our proposed targets and actions will make sense.

**Assumption 2.** *There will be no great technology revolution bringing revolutionary alternative renewable energy.*

There is no sign for the emergence of some new and revolutionary alternative energy up to now, so this assumption can help to eliminate the effects of the small probability event.

**Assumption 3.** *No destructive catastrophe (such as great earthquake) and no human-made disasters (such as war) will happen by 2050.*

These kinds of disasters happened rarely in the history record, so we neglect their effects in our model.

**Assumption 4.** *The provided data is realistic and accurate to a certain degree.*

Despite the incompleteness of the data and some tolerant error in statistics, we make this assumption to guarantee one valid solution.

Table 1: Abbreviation of Relevant Terms

| Abbreviation | Full Name |
|:---:|:---:|
| RC | Renewable Consumption |
| UC | Unrenewable Consumption |
| RP | Renewable Production |
| UP | Unrenewable Production |
| RI | Renewable Import |
| RE | Renewable Export |
| RED | Renewable Expenditure |
| UED | Unrenewable Expenditure |
| CPC | Consumption Per Capita |
| EPC | Expenditure Per Capita |
| RPES | Renewable Production Entropy by Source |
| UPES | Unrenewable Production Entropy by Source |
| RCES | Renewable Consumption Entropy by Source |
| UCES | Unrenewable Production Entropy by Source |
| SRP | Share of Renewable Production |
| SRC | Share of Renewable Consumption |
| RCPR | Renewable Consumption-Production Ratio |
| TCPR | Total Consumption-Production Ratio |
| RCEES | Renewable Consumption Entropy by End-use Sector |
| UCEES | Unrenewable Consumption Entropy by End-use Sector |

# 3 Abbreviation and Definitions

For compactness, we define a series of abbreviations for some notions concerning energy profile in Table 1.

# 4 CAFE: Characterization, Analysis, Forecasting, and Evaluation of Energy Profile

In this section, we firstly create the energy profile based on provided data. Then we propose a novel model: CAFE, to comprehensively characterize, analyze, forecast, and evaluate the energy profiles of CA, AZ, NM, and TX, based on the provided data set comprised of 605 variables from 1960 to 2009 [5].

## 4.1 Energy Profile Formulation

First of all, we formulate an energy profile for each state. The energy profile should present a panoramic view of energy use. We macroscopically divide energy into *renewable* (electricity, ethanol fuel, nuclear fuel, wood etc.) and *unrenewable* (oil/gasoline/petroleum, coal, natural gas) ones. We extract items reflecting the global information about different

aspects of energy in a state, mainly in terms of import, export, production, consumption, and expenditure. Specifically, we select the following items which possess the global property: RC, UC, RP, UP, RI, RE, RED, UED, CPC, EPC.
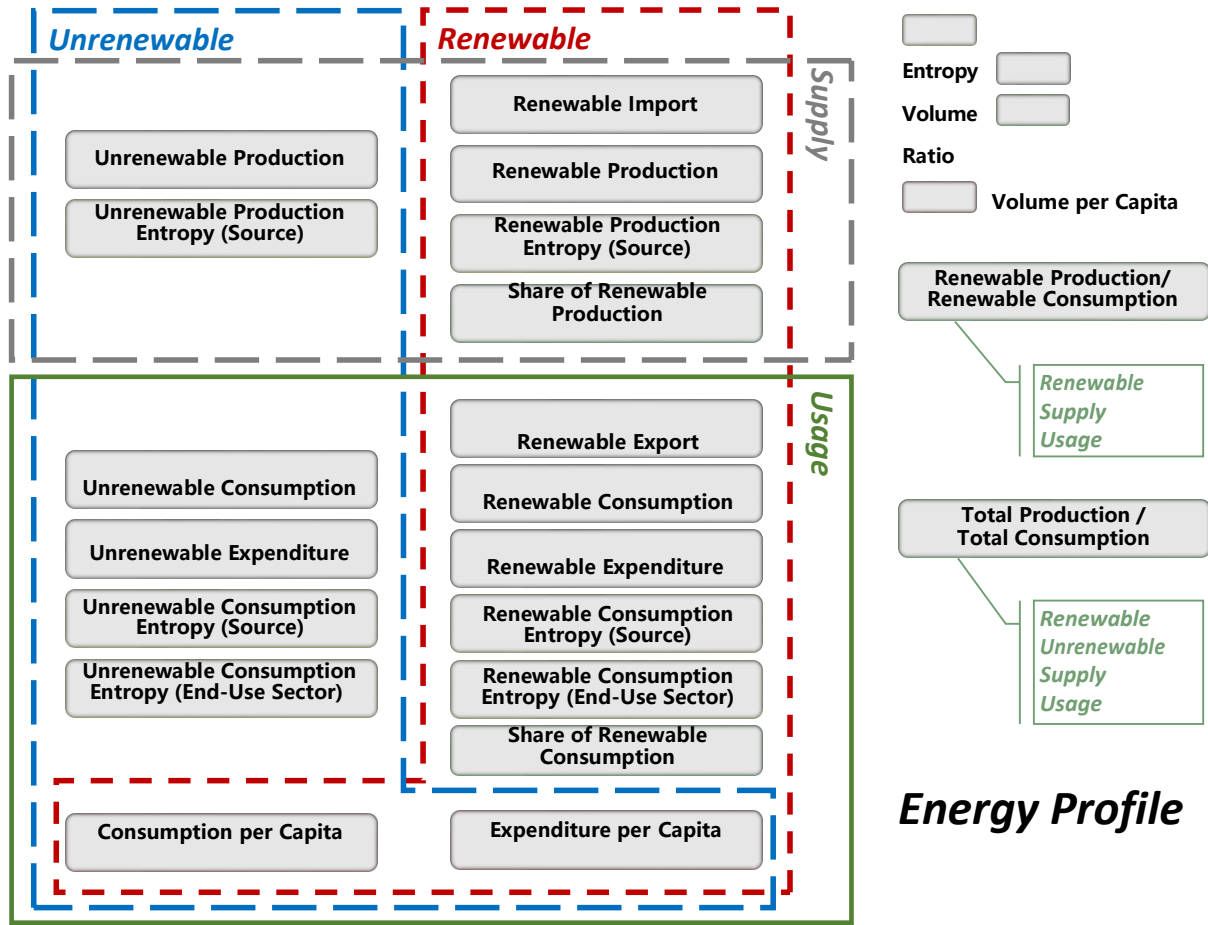


Figure 2: The Energy Profile of a State

To explore the underneath information within these items, we make some aggregations to form new items listed in Table 2. Inside, RPES/UPES/RCES/UCES (RPES as an epitome) characterizes energy usage distribution over different kinds of sources while RCEES/UCEES (RCEES as an epitome) shows that over different sectors, serving as a supplement to items with totality property; SRP/SRC represents the popularity of renewable energy use; RCPR/ TCPR evinces the balance of energy use in a state. All these aggregated items enrich our consideration in energy profile characterization.

Table 2: Aggregated Items in Energy Profile

| Item | Definition |
|------|------------|
| RPES/RCEES | $\sum_i p_j \log(1/p_j)$ [3] |
| SRP | RP/(RP+UP) |
| SRC | RC/(RC+UC) |
| RCPR | RC/RP |
| TCPR | (RC+UC)/(RP+UP) |

---

[3] $p_j$ is the corresponding portion of $j$-th (un)renewable energy in sources/sectors.

Then combining the aggregated items with selected original items, we propose a taxonomy of items in energy profile as Figure 2, to help readers understand the relationship among extracted items. In Figure 2, we group these items into four sub-classes based on whether the energy is renewable and supply or usage of the energy. Moreover, these items can be classified into four categories with different physical dimensions: entropy, volume, ratio, and volume per capita. Items concerning volume show the concrete values of different energy aspects, while those in the relative dimension present the relationship between different aspects of energy. By this classification, we clarify the relationship of items in energy profile in a logical and hierarchical way.

## 4.2 Energy Profile Characterization and Analysis

### 4.2.1 GPR model: Evolution of Energy Profile

To help us understand the evolving pattern of the energy profile, we probe into a generative model to characterize the historical evolution. We first make an observation on the given data to form an intuitive understanding of the longitudinal data. We select four typical time series among extracted items, which present different shapes and trends, as shown in Fig. 3. Specifically, renewable expenditure of Arizona (Fig. 3(a)) generally presents an ascending trend with tiny fluctuations; renewable export of Arizona (Fig. 3(b)) has a sudden increase while renewable consumption entropy by source (Fig. 3(c)) appears to fluctuate near a certain value.

Based on the observations, we summarize the following challenges in characterization of these time series:

· The fitting model should characterize the basic evolving trend of time series;

· The strong fluctuations of item values requires a model to capture the instability;

· The model should accommodate the random noise existing in time series.

In view of these challenges, we adopt the Gaussian Process Regression (GPR), an non-parametric probabilistic model, to fit the time series. The model of GPR can be generalized as follows. Assume that noisy data

$$y(x) = f(x) + s(x) \tag{1}$$
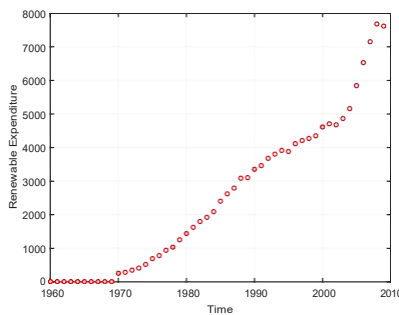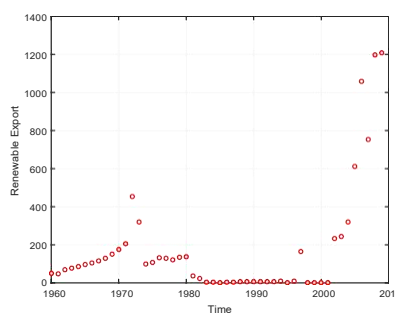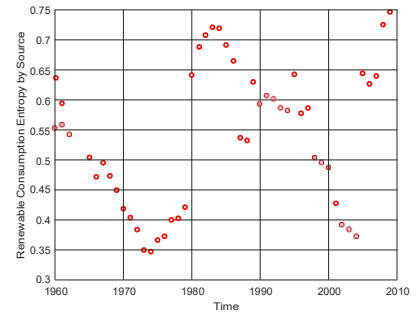


(a) RED of Arizona.          (b) RE of Arizona.          (c) RCES of Arizona.

Figure 3: Three typical time series of longitudinal energy profile data.

where $f(x)$ is the ground-truth value and $\epsilon(x) \sim \mathbf{N}(0, \sigma^2)$ is the temporally irrelevant random noise. GPR model assumes $f(x)$ to satisfy the Gaussian process, i.e.,

$$f(x) \sim GP(\mu_x, K_f), \tag{2}$$

where $\mu_x$ is the mean value and $K_f$ is the covariance matrix called kernel. A natural selection for $K_f = (k_f(x_i, x_j))_{n \times n}$ is the Gaussian function:

$$k_f(x_i, x_j) = c_0 \exp\left(-\frac{(x_i - x_j)^2}{2\sigma_0^2}\right), \tag{3}$$

where $c_0$ is a constant and $\sigma_0$ represents the decaying rate of auto-correlation.

Consider a time series $\{y(x) \in \mathbf{R}\}$ where $x$ denotes a time point and $y$ is the value at $x$. Suppose that $\mathbf{y} \in \mathbf{R}^{n \times 1}$ constitutes the $n$ observed time series value, then $\mathbf{y} \sim GP(\mu, K)$. Here $\mu$ is the mean value vector and

$$K = (k(x_i, x_j))_{n \times n} = K_f + \sigma^2 I, \tag{4}$$

where $\sigma$ reflects the noise level. For a new value $x^*$, predict its corresponding $y^*$. We can easily find that

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu^* \end{bmatrix}, \begin{bmatrix} K_0 & K_1^T \\ K_1 & K_2 \end{bmatrix}\right), \tag{5}$$

where $K_0 = K$, $K_1 = [k(x^*, x_1), k(x^*, x_2), \cdots, k(x^*, x_n)]$ and $K_2 = k(x^*, x^*)$. Therefore we can obtain the conditional distribution of $y^*$ over $\mathbf{y}$,

$$y^*|\mathbf{y} \sim N(K_1 K_0^{-1} \mathbf{y}, K_2 - K_1 K_0^{-1} K_1^T). \tag{6}$$

Hence our predicted $y^*$ satifies $y^* \sim \mathbf{N}(\mu_*, \sigma_*^2)$ where,

$$\mu_* = K_1 K_0^{-1} \mathbf{y}, \quad \sigma_*^2 = K_2 - K_1 K_0^{-1} K_1^T + \sigma. \tag{7}$$

**Justification of GPR model:**

- $\beta = \|K_0^{-1}\|_\infty$ reflects the basic evolving trend of time series.

- $\alpha = \frac{1}{n} \sum \sigma_*^2$ quantitatively characterizes the fluctuation of the time series.

- $\sigma^2$ captures the random noise level.

GPR model makes a probabilistic prediction of outputs with confidence intervals in consideration of the stochastic errors. Furthermore, GPR renders higher correlation for temporally closer values, coinciding with the fact that closer information is of greater significance in prediction of energy profile.

By adopting GPR, we fit the time series in Figure 3 as Figure 4.

Since GPR does not output a deterministic value, we additionally show the 95%-confident intervals of each output. The gap surrounded by the upper and lower confident lines is called *confidence interval*. As we can see, for RED of Arizona with a consistent increasing trend, the confident interval is fairly small, showing the **high stability and predicting reliability** of its evolution. By contrast, for RCES of Arizona with the greatest fluctuation, the

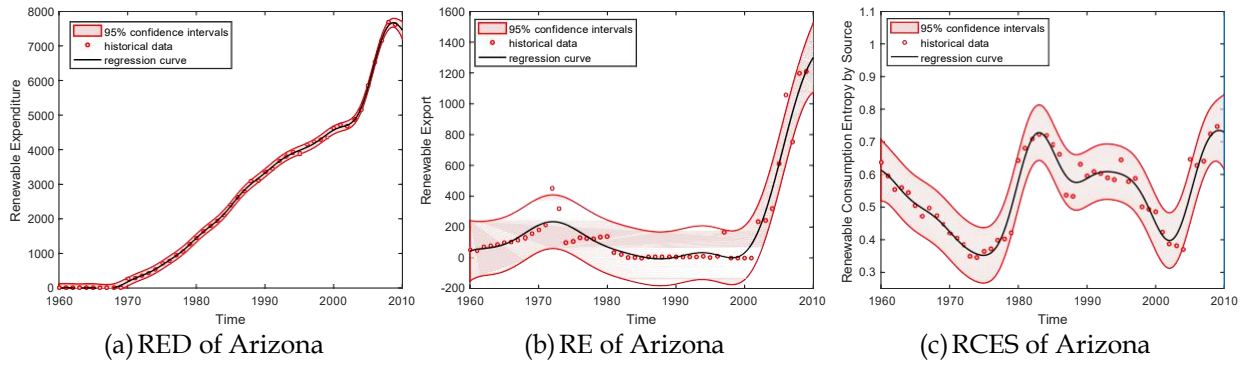(a) RED of Arizona      (b) RE of Arizona      (c) RCES of Arizona

Figure 4: GPR curve of three typical time series

confident interval is large, which indicates that the evolution of renewable expenditure is of **great randomness, low predictability, and high sensitivity to outer influential factors**.

We further apply our model into the remaining energy profile time series and present the regression parameters of AZ in Table 3 and the corresponding regression curve in Fig. 5 (the placing order is the same as Table 3). As we can see, $\beta$ reflects the general trend of the evolution, $a$ represents the temporal fluctuation of measurements while $\sigma^2$ shows the noise error. Items (RC, UC, etc.) have $a$ $\sigma^2$, which means that fluctuation takes the predominant part while random noise can be neglected; For those where $a$ is close to $\sigma^2$ (UCES, SRC, etc.), it is not reasonable to omit the impact of random noises.

Table 3: GPR regression results for energy profiles of Arizona (1960-2009) (The arrow directions represent the basic evolving trends while L and S is the abbreviation of large and small respectively)

|  | RC | UC | RP | UP | RI | RE | RED | UED | CPC | EPC |
|---|---|---|---|---|---|---|---|---|---|---|
| $\beta$ | 90679 | 782873 | 88454 | 298941 | 124 | 297 | 2657 | 3400 | 0.11 | 1328 |
| Trend | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | → | ↑ |
| $a$ | 10164 | 1.89 | 10161 | 3.33 | 190 | 82 | 52 | 4.8 | 5205 | 98 |
| Fluctuation | L | S | L | S | S | S | S | S | L | S |
| $\sigma^2$ | 0.13 | 0.03 | 0.13 | 0.02 | 0.06 | 0.05 | 0.02 | 0.06 | 0.22 | 0.09 |
| Noise | L | S | L | S | S | S | S | S | L | S |
| *MAPE* | 0.05 | 0.02 | 0.05 | 0.63 | 0.23 | 0.35 | 0.17 | 0.08 | 0.01 | 0.02 |
|  | RPES | UPES | RCES | UCES | RCEES | UCEES | SRP | SRC | RCPR | TCPR |
| $\beta$ | 0.56 | -0.23 | 0.56 | 2.40 | -0.42 | -0.70 | -0.98 | 0.32 | 1.60 | 2.11 |
| Trend | ↑ | ↓ | → | ↑ | ↓ | ↓ | ↓ | ↑ | ↑ | ↓ |
| $a$ | 2234 | 0.17 | 412 | 710 | 458 | 1290 | 0.01 | 0.04 | 231 | 980 |
| Fluctuation | L | S | L | L | L | L | S | S | L | L |
| $\sigma^2$ | 0.34 | 0.07 | 0.28 | 0.31 | 0.02 | 0.34 | 0.01 | 0.11 | 0.31 | 0.01 |
| Noise | L | S | L | L | S | L | S | S | L | S |
| *MAPE* | 0.02 | 0.25 | 0.04 | 0.01 | 0.07 | 0.05 | 0.01 | 0.08 | 0.01 | 0.01 |

**Error Analysis:** Table 3 also shows the Mean Absolute Percent Error (MAPE) between regression values and ground-truth values. The low MAPEs corroborate the regression precision of GPR model in fitting energy profile time series.
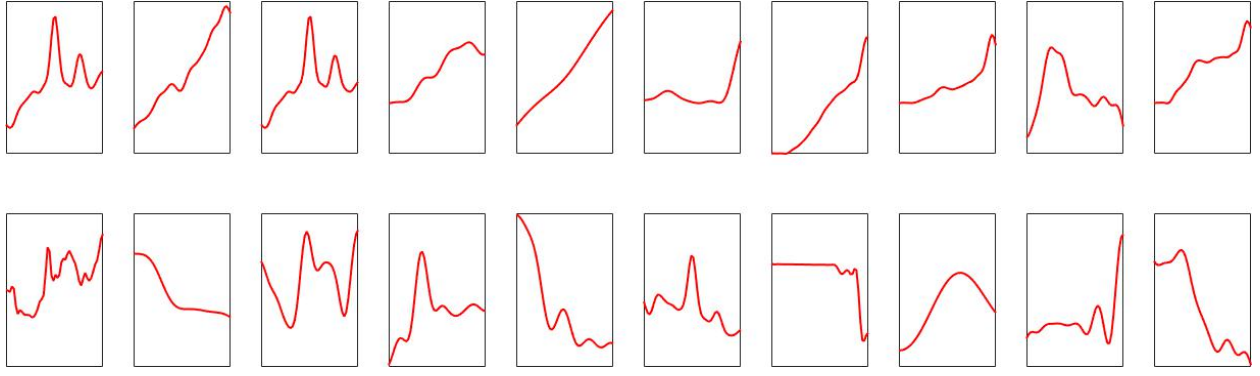
Figure 5: Evolution of Energy Profile for AZ (the placing order is the same as Table 3)

### 4.2.2  Similarity Analysis of Energy Profile

We take a step further to study the similarity of energy profile evolution among four states. Shown in Figure 4, the temporal fluctuation entangles accurate analysis of the relationship among four states over each item, due to the uncertain time-variation of each time series and our incomplete knowledge of other potential influential factors. For such an unclear situation, we adopt the Gray Relational Analysis (GRA) to unveil the relationship among energy profile in four states. GRA performs reliably under systems with information incompleteness and uncertainty, and with complicated multivariate interrelationships [6, 7], by pursuing an appropriate result acceptable in real cases instead of persisting in an optimum [8].

We conduct our GRA-based analysis as follows: For item $e$, suppose we consider the values of it for $T$ years, then we can obtain the time series $y_i = (y_i(1), y_i(2), ..., y_i(T))$ for state $i$. Then for two states $i$ and $j$, we define $\xi_{ij}$, which represents the relationship between *comparability vector* $y_i$ and *reference vector* $y_j$. Specifically,

$$\xi_{ij} = \frac{\min\limits_{t\in\{1,2,...,T\}} \min\limits_{i\neq j} |y_i(t) - y_j(t)| + \max\limits_{t\in\{1,2,...,T\}} \max\limits_{i\neq j} |y_i(t) - y_j(t)|}{|y_i(t) - y_j(t)| + \max\limits_{t\in\{1,2,...,T\}} \max\limits_{i\neq j} |y_i(t) - y_j(t)|}. \tag{8}$$

Then the gray relational coefficient (GRC) is

$$\rho_{ij} = \frac{\sum_{k=1}^{T} \xi_{ij}(k)}{T}, \tag{9}$$

which characterizes the correlation.

Nevertheless, conducting GRA only measures the similarity on totality property of each item, as shown in Eqn. (8), which cannot characterize the similarity on evolution tendency. In real cases, however, tendency similarity counts much as one can characterize the future trend of a profile more precisely. To tackle it, we harness the *Kendall Rank Correlation Coefficient (KRCC)* to model the tendency similarity. KRCC considers two random arrays $y_i = (y_i(1), y_i(2), ..., y_i(T))$ and $y_j = (y_j(1), y_j(2), ..., y_j(T))$. For a pair of observations $(y_i(t_1), y_j(t_1))$ and $(y_i(t_2), y_j(t_2))$ $(t_1 \neq t_2)$, if $(y_i(t_1) - y_i(t_2))(y_j(t_1) - y_j(t_2)) > 0$ we call the pair is *concordant*, while if $(y_i(t_1) - y_i(t_2))(y_j(t_1) - y_j(t_2)) < 0$ means *discordant*. KRCC is then defined as

$$\tau(X, Y) = \frac{\{N_c\} - \{N_d\}}{T(T-1)/2}. \tag{10}$$

$$\xi_{ij} = \frac{\displaystyle\min_{t\in\{1,2,\dots,T\}}\ \min_{i\neq j}|y_i(t) - y_j(t)| + \max_{t\in\{1,2,\dots,T\}}\ \max_{i\neq j}|y_i(t) - y_j(t)|}{|y_i(t) - y_j(t)| + \displaystyle\max_{t\in\{1,2,\dots,T\}}\ \max_{i\neq j}|y_i(t) - y_j(t)|}. \tag{8}$$

Then the gray relational coefficient (GRC) is

$$\rho_{ij} = \frac{\sum_{k=1}^{T}\xi_{ij}(k)}{T}, \tag{9}$$

which characterizes the correlation.

Nevertheless, conducting GRA only measures the similarity on totality property of each item, as shown in Eqn. (8), which cannot characterize the similarity on evolution tendency. In real cases, however, tendency similarity counts much as one can characterize the future trend of a profile more precisely. To tackle it, we harness the *Kendall Rank Correlation Coefficient (KRCC)* to model the tendency similarity. KRCC considers two random arrays $y_i = (y_i(1), y_i(2), \dots, y_i(T))$ and $y_j = (y_j(1), y_j(2), \dots, y_j(T))$. For a pair of observations $(y_i(t_1), y_j(t_1))$ and $(y_i(t_2), y_j(t_2))$ $(t_1 \neq t_2)$, if $(y_i(t_1) - y_i(t_2))(y_j(t_1) - y_j(t_2)) > 0$ we call the pair is *concordant*, while if $(y_i(t_1) - y_i(t_2))(y_j(t_1) - y_j(t_2)) < 0$ means *discordant*. KRCC is then defined as

$$\tau(X, Y) = \frac{\{N_c\} - \{N_d\}}{T(T-1)/2}. \tag{10}$$

Then we can obtain the integrated KRCC denoted as *IKRCC$_{ij}$* in the same way.



(a) GRC　　　　　　　　　　　　　　　　　　　(b) KRCC

Figure 6: Heat map for GRC and KRCC

## 4.3　Influential Factors of Similarities

To understand the profile similarities and differences much in depth, we then explore the impacts of other possible influential factors including geography, climate, industry and population. Note that geography and climate can be viewed as constants while industry and population are evolving temporally. Therefore, for a constant factor *G* we use *Absolute Mean Error* between two states *i* and *j*, denoted as

$$\gamma_{ij}^G = \frac{|G_i - G_j|}{|G_i|},\qquad(13)$$

to characterize the interstate similarity of factor $G$. For an evolving factor $H$ we use the *Pearson Correlation Coefficient (PCC)* of it in two states $i$ and $j$, whose definition is

$$\rho_{ij}^H = \frac{\sum_k (H_k^i - \bar{H}^i)(H_k^j - \bar{H}^j)}{\sqrt{\sum_l (H_l^i - \bar{H}^i)^2 \sum_l (H_l^j - \bar{H}^j)^2}},\qquad(14)$$

where $H_i$ and $H_j$ are the arrays of $H$.

To study the influential factors, our empirical intuition is: if two states with a similar factor (like geography) share the similar energy profile, then we treat this factor is related to the interstate similarity of energy profile. In Figure 7, we plot the factor similarity v.s. *IGRC* for $2C_4^2 = 12$ state pairs. As we can see, the scattered points appear to have a consistent trend, which indicates that the factor similarity is related to *IGRC*.

Then we probe into the influential factors for *IKRCC*. We adopt Pearson Correlation Coefficient (PCC) and Partial Relational Coefficient (PRC) to study the correlation. The advantage of PRC over PCC is that PRC takes the interdependence among different influential



(a) Area

(b) Temperature
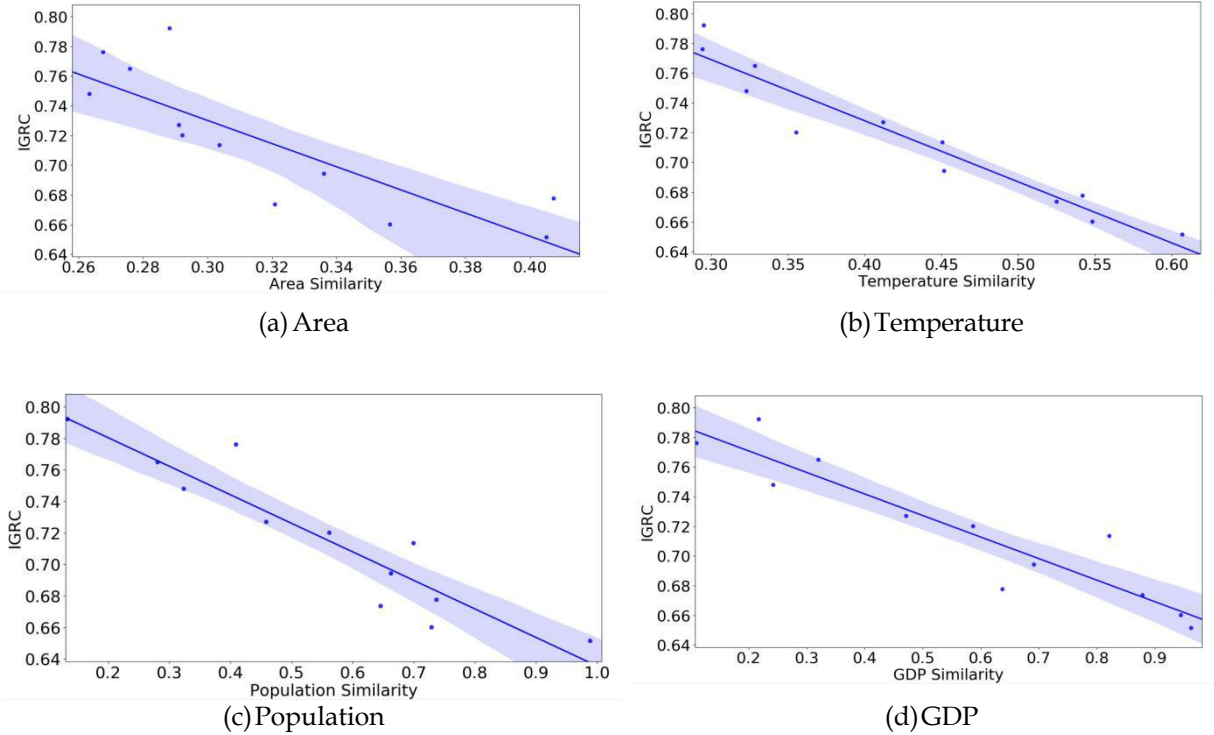
(c) Population

(d) GDP

Figure 7: Factor Similarity v.s. IGRC

factors into considerations and capture the correlation excluding the impact from other factors. Table 4 shows the PCC and PRC between influential factors and *IKRCC*: PCC shows that all of the potential criteria have a negative correlation within the two states involved.

Table 4: Pearson Correlation Coefficient and Partial Relational Coefficient.

|        | Area        | Water Percentage | Altitude | Rainfall       |
|--------|-------------|------------------|----------|----------------|
| **PCC** | -0.45       | -0.26            | -0.39    | -0.44          |
| **PRC** | -0.57       | -0.14            | 0.27     | 0.34           |
|        | Temperature | Population       | GDP      | GDP Per Capita |
| **PCC** | -0.61       | -0.28            | -0.21    | -0.42          |
| **PRC** | -0.71       | -0.72            | 0.80     | -0.37          |

## 4.4   Energy Profile Evaluation

Following the energy profile characterization, we will analyze which state turns out to have the best profile in terms of renewable energy usage. A good profile for renewable energy should be integratedly judged and ultimately indicate a positive tendency for its greater influence in society. In our work, we designate seven criteria based on different angles to measure how good a profile is in regards to renewable energy use, list in detail in Table 5. New criteria include: RCPC (Renewable Consumption Per Capita), IRRC (Increasing Rate of Renewable Consumption), and REGR (Renewable Expenditure-GDP Ratio). The proposed criteria synthetizes popularity, diversity, and economic potentials, ensuring higher social impact of renewable energy under a better energy profile for renewable energy use.

Table 5: Criteria for Evaluating Energy Profile

| Abbreviation | Significance towards RE | Entropy weight |
|--------------|-------------------------|----------------|
| RCES  | Consumption Variety    | 0.183 |
| RCEES | Consumption Variety    | 0.191 |
| RCPR  | Economic Balance       | 0.105 |
| SRC   | Consumption Proportion | 0.120 |
| RCPC  | Individual Consumption | 0.189 |
| IRRC  | Consumption Trend      | 0.107 |
| REGR  | Expenditure Proportion | 0.101 |

Judging the best profile requires reasonably melting the above seven criteria into a synthetic one. Here we again adopt *Entropy Weight Method* to allocate weights to each criterion. In view of our objective to selecting the *best* profile, we utilize the TOPSIS[2] method, whose core is to directly identify the best profile by minimizing the distance to the 'virtual optimum' where all criteria reach the optimum. To implement the method, we first normalize each criterion value to eliminate the problem of different scales. Use $Z_{ij}$ to denote the value of the *j*-th criterion for state *i* and its corresponding normalized value $Z^0_{ij}$ can be calculated

---

[2]TOPSIS: Technique for Order Preference by Similarity to an Ideal Solution

by

$$Z_{ij}^0 = \frac{Z_{ij} - \min_i Z_{ij}}{\max_i Z_{ij} - \min_i Z_{ij}}. \tag{15}$$

Second, we calculate the distance to the "virtual optimum",

$$D_i^+ = \sqrt{\sum_{j=1}^{7} W_j \cdot (Z_{ij} - Z_j^+)^2}, \tag{16}$$

$$D_i^- = \sqrt{\sum_{j=1}^{7} W_j \cdot (Z_{ij} - Z_j^-)^2}, \tag{17}$$

where $Z_j^+$, $Z_j^-$ stand for the best and worst value of the $j$-th criterion among four states; $W_j$ is the entropy weight for the $j$-th criterion. Finally, we can derive the evaluation score for state $i$,

$$Score(i) = \frac{D_i^-}{D_i^+ + D_i^-}, \tag{18}$$

The evaluation results are presented in Table 6, whose first seven columns are the normalized criteria, and the last column is the score for each state. As we can see, CA ranks the first in our final criterion, with TX following behind, while AZ ranks the third and NM is at the bottom. This matches our sense that CA and TX are more developed, and they should go further in renewable energy use compared with less developed AZ and NM.

Table 6: Evaluation Results of Renewable Energy Usage for Four States

|     | RCES | RCEES | RCPR | SRC  | RCPC | IRRC | REGR | Score |
|-----|------|-------|------|------|------|------|------|-------|
| AZ  | 0.16 | 0     | 0.95 | 0.69 | 0.16 | 0    | 1    | 0.434 |
| CA  | 1    | 1     | 0.56 | 1    | 0    | 1    | 0    | 0.628 |
| NM  | 0.32 | 0.21  | 0    | 0.38 | 0.29 | 0.52 | 0.64 | 0.417 |
| TX  | 0    | 0.21  | 1    | 0    | 1    | 0.83 | 0.88 | 0.534 |

## 4.5　Energy Profile Prediction

We now make predictions on each energy profile of the four states. Following the GPR model in profile characterization (Section 4.1), we first make prediction by GPR, results of which are presented in Figure 8. We can observe that all items turn into a stable increase after 10 years. This is because in long-term perspective, the prediction should reflect the general trend, and become more 'conservative' especially in predicting possible fluctuations since uncertainty increases. Therefore, for those showing a general increasing trend, like RED and RE, the long-term prediction shows a stable uphill, while for those showing violent fluctuation like RCES, the prediction only turns a slight increase. This phenomenon echos our characterization of item value evolution (Section 4.1): the higher predictability the evolution is, the less conservative the long-term prediction will be.

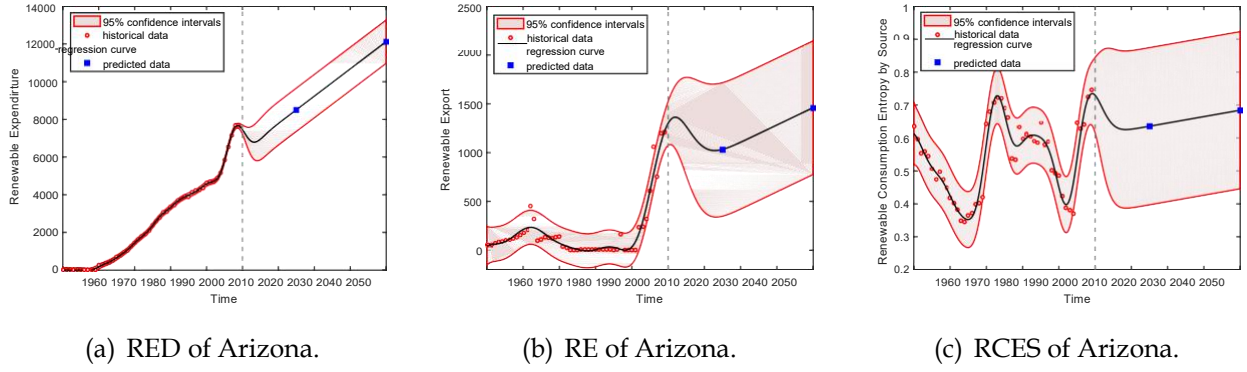(a) RED of Arizona.     (b) RE of Arizona.     (c) RCES of Arizona.

Figure 8: GPR-Based Energy Profile Prediction of Arizona.

However, although GPR presents a global prediction, it fails to focus on the local information, thus making unreasonable short-term prediction. For example, for RE of Arizona (Figure 8(b)), we can discover that within a short period after 2009, there is a counterintuitive decrease opposite to the sharp increase within several years before 2009. This is because GPR synthesizes all the historical data, and thus drags down the sudden increase after 2009.

To refine the short-term prediction, a local prediction model should be taken. We design two novel techniques to refine our model.

**Technique 1: Sliding Window Mechanism**

The *Autoregressive Moving Average Model (ARMA)* model is an autoregressive predicting method with priority in short-term prediction [9]. Its general representation is ARMA(p,q) as follows:

$$y(t) = c + \epsilon(t) + \sum_{i=1}^{p} \phi_i y(t-i) + \sum_{i=1}^{q} \theta_i \epsilon(t-i), \qquad (19)$$

where $s(t)$ is the white noise at time $t$, while $\varphi_i$ and $\theta_i$ are both weights. The sliding window
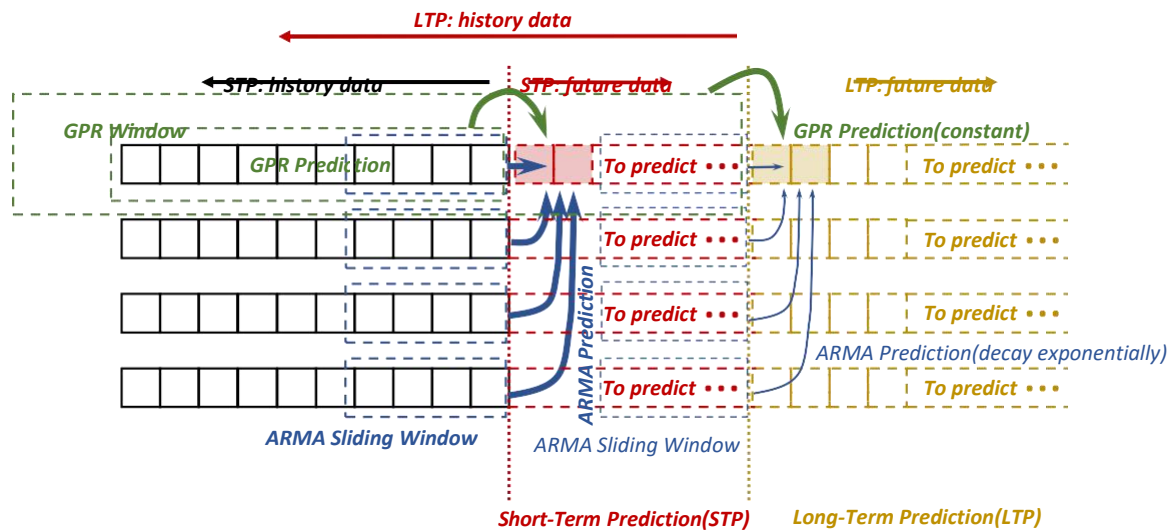


Figure 9: ARMA-GPR Hybrid Model in short-term and long-term prediction

mechanism, which restricts our consideration within a small temporal range near the current

time, keeps away from the 'outmoded' data of less contribution and even bringing harmful offset in short-term prediction. Specifically, we set the current time $t$ and only consider time interval $[t \ k + 1, t]$, where $k \ t$. An intuitive example is illustrated in Figure 9. Moreover, as we target at following the trend around current time in short-term prediction with higher accuracy, we need to precisely characterize how much the item value changes, that is to apply ARMA model on the increments of item values, i.e.,

$$\Delta y(t) = c + \epsilon(t) + \sum_{i=1}^{p} \phi_i \Delta y(t - i) + \sum_{i=1}^{q} \theta_i \epsilon(t - i). \tag{20}$$

### Technique 2: 'Look Aside' Latitudinal Refinement

As the existing similarities of profiles among four states, discussed in Section 4.2, it is inadequate to solely take longitudinal study on the profile evolution of a single state. Instead, latitudinal perspectives should be embedded, i.e., taking the correlation of profiles among states into the prediction of a specific state. The intuitive illustration is shown in Figure 9. To quantify the integration of interstate correlation, we add a weight $\delta$ to the predictions by ARMA in other states, specifically,

$$\Delta \tilde{y}_i(t) = b \Delta y_i(t) + (1 - b) \sum_{j \neq i} \Delta y_j(t). \tag{21}$$

After obtaining the integrated increment $\Delta \tilde{y}_i(t)$, we predict the item value by

$$y^A(t) = y(t - 1) + \Delta \tilde{y}_i(t). \tag{22}$$

### Incorporation of Techniques

We now combine the prediction by ARMA and GRP and form a novel hybrid model, which we call ARMA-GPR model. Note that from short-term to long-term prediction, the local factor counts less while global factor counts more. Hence, we render an exponentially decaying weight $\kappa(t)$ to the local prediction. Concretely, the prediction at time $t$ should be

$$y^*(t) = \kappa(t)y^A(t) + (1 - \beta(t))y^G(t), \tag{23}$$

where $y^G(t)$ denotes the prediction given by GPR and $\kappa(t) = \exp(-a_0(t - t_0))$ ($c_0 > 0$ and in this problem $t_0 = 2009$).

We plot the prediction results of ARMA-GPR model in Fig. 10. Comparing with the prediction of GPR, the predicted value given by ARMA-GPR model appears to be 'higher' than GPR predicted value. This is because the time series before 2009 present a ascending trend, the sliding window of the ARMA model would capture this trend and 'pull' up the predicted value. With the time going by, the ARMA-GPR predicted value inclines to converge to the GPR predicted value, which results from the time-decaying weight.

# 5    Targets and Actions for Interstate Energy Compact

In this section, we put our CAFE model for energy profile into practical policies. Specifically, we determine quantified goals for renewable energy use for 2025 and 2050, and propose specific measures for accomplishing such goals by the four states.
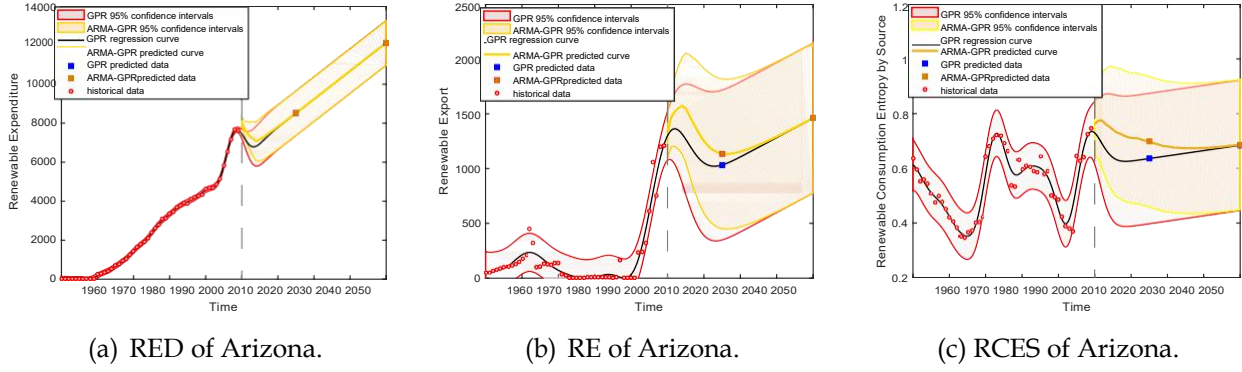
(a) RED of Arizona.  (b) RE of Arizona.  (c) RCES of Arizona.

Figure 10: ARMA-GPR-Based Energy Profile Prediction of Arizona.

## 5.1 Targets: Total Renewable Consumption Based Optimization

Based on our observation on provided data, we identify that the *renewable consumption (RC)* accounts for a dominant proportion of total renewable energy use. The predominance of RC enlightens us to concentrate on consumption issues in drafting the targets while putting aside others like RP, RI, and RE.

Before we get down to setting targets for energy usage, we make some further assumptions based on the general ones in Section 2.

**Assumption 5.** *The ultimate aim of the compact is to maximize the global benefits over states.*

**Assumption 6.** *The governors of each state incline to be 'selfish' and , that is to say, they are unwilling to sacrifice their own benefits in the consideration of the compact.*

Based on Assumption 5, the goal of the compact should be maximizing total renewable consumption (TRP) among the states. However, solely considering the amount of consumption covers the renewable energy use partially, as analyzed in Section 4.3. Therefore, a more reasonable way to deciding goals should stem from our proposed criteria evaluating the energy profile (Section 4.3). Hence the target setting problem can be formulated as a multi-objective programming. To integrate the evaluation (i.e., objective function) of four states, we define *Total Renewable Consumption Profile (TRCP)* and

$$TRCP = \sum_{i=1}^{4} b_i \tilde{Score}_p(i), \tag{24}$$

where $\tilde{Score}_p(i)$ is the evaluation score after implementation of compact.

When determining $b_i$, we should notice the common sense that the interstate compact ought not to further expand the gap of renewable energy use among these states, causing more serious economic disparity. Therefore, an intuitive way is to offset the 'dominance' of states with higher profile score in $TRCP$ by setting proper $w_i$. Specifically we set $w_i = 1/Score(i)$, where $Score(i)$ is the current profile score of state $i$. Note that the maximization of $TRCP$ would only concern three criteria related to RC: SRC, RCPR, and REGR.

Moreover, there are two restrictions for the states during their chase for TRC maximization:

- **RC Satisfaction:** The compact should be satisfied by all four states, hence it should not cripple the consistent economic development of any state.

- **RC Limitation:** The TRC should not exceed total renewable production (TRP) in every year within the four states.

For the first restriction, we suppose that the satisfactory lower bound of RC in each state for one year is the *RC prediction* given in Section 4.4. Mathematically, let RC prediction be $\tilde{RC}$, therefore $RC_i \geq \tilde{RC}_i, i = 1, 2, 3, 4$; For the second restriction, we need to ensure $\sum_{i=1}^4 RC_i \leq \sum_{i=1}^4 RP_i$. Combining our objective and restrictions, we formulate the following basic TRCP maximization problem:

$$\max_{RC_i, i=1,2,3,4} TRCP = \sum_{i=1}^4 \frac{\tilde{Score}_p(i)}{Score(i)}$$

$$s.t. \quad \begin{cases} RC_i \geq \tilde{RC}_i & (i = 1, 2, 3, 4), \\ \sum_{i=1}^4 RC_i \leq \sum_{i=1}^4 RP_i. \end{cases}$$

Apparently, the problem is a linear programming, thus easy to solve. However, we observe that there are two limitations in the original TRCP maximization problem:

. **Inhomogenity:** The distribution of *RC* among states is heavily skewed. Most of the *RC* enhancement is offered to the New Mexico, while the other three states only enjoy tiny increase. This extreme inhomogeneity inhibits the RC development in states other than NM, deviating from the mutuality of the compact.

. **Infeasibility:** It is possible that no feasible solution exists in maximizing *TRCP* due to the hard constraints on *RC*. For example when *RP* is smaller than our predicted $\tilde{RC}$, then chances are that the feasible region is an empty set.

To this end, we modify the original optimization problem into the following renovated version,

$$\max_{RC_i, i=1,2,3,4} TRCP' = \sum_{i=1}^4 \frac{\tilde{Score}_p(i)}{Score_i} - \lambda \left( \sum_{i=1}^4 RC_i - \sum_{i=1}^4 RP_i \right)^2$$

$$s.t. \quad RC_i \geq d_i \cdot \tilde{RC}_i \quad (i = 1, 2, 3, 4),$$

where $\lambda$ is positive predefined parameters and $d_i (i = 1, 2, 3, 4)$ are the expected benefit of each state, which can be set by the governors.

Intuitively, there are two adjustments in the new optimization problem. The first is introducing an additive RC requirement *i.* It tightens the restriction of RC enhancement, thus ensuring an acceptable increment in each state, which dissolves the **inhomogenity** limitation. The second is converting the *RC Limitation* constraint into a penalized term in the objective function, where $\lambda$ is the penalized factor. The penalized term restricts small discrepancy between TRC and TRP which meets the balance requirement in practice. Meanwhile, it relaxes the hard requirements that $RC \leq RP$ based on the incompleteness of the

provided data, and the possible renewable energy transformation with states other than the four mentioned ones, which dissolves the **infeasibility** limitation.

Then we apply our model to setting of targets in 2025 and 2050. Table 7 shows the renewable consumption targets as well as improvement percentage compared with the predicted RC for 2025 and 2050. Here we consider $d_i$ = 10% for 2025, $d_i$ = 20% for 2050, $i$ = 1, 2, 3, 4, and set = 0.015. We will further discuss the setting of in sensitivity analysis.

Table 7: Renewable Consumption Targets for 2025 and 2050 ($10^{14}$Btu)

| Year | AZ | CA | NM | TX |
|------|-----|------|-----|-----|
| **2025** | 1.63(24%) | 10.90(10%) | 0.54(81%) | 3.12(10%) |
| **2050** | 2.10(58%) | 14.73(20%) | 0.77(113%) | 4.45(20%) |

Then we study the effects of our proposed targets on energy profiles. Specifically, we compare the predicted values of RCPR, SRC, and RCPC with the proposed targets and those without targets. The results are shown in Figure 11. We notice that for all three items, the predicted values with targets all outstrip those without targets in both 2025 and 2050, showing the probable promotion brought about by the targets.
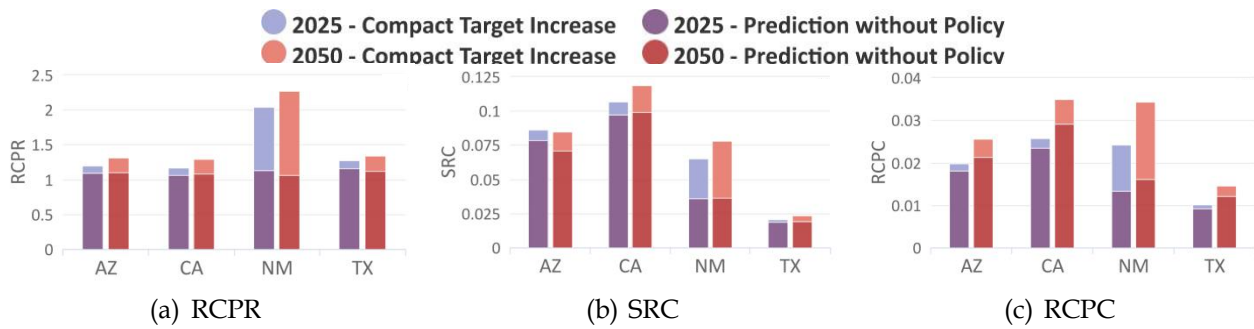


Figure 11: The Effects of Proposed Policies in Energy Profile Enhancement.

Moreover, we also calculate the evaluation criteria based on the RC targets and present the results in Fig. 12. As we can see, SRC, RCPR, and, REGR for four states achieve significant improvement under the RC targets.

## 5.2 Actions: Motivating Renewable Energy Use in NM, TX, and AZ

**Principles for Actions:** (i) Since our goals are predicated on the renewable energy consumption (RC), the actions should first contribute to the achievement of RC requirements;



Figure 12: Radar Map: Energy Profile Evaluation and Target

(ii) If RC requirements are fulfilled, the actions should also bring about positive effects on related items, for example the consumption entropy and the share of renewable consumption;
(iii) Actions should also be prediction-based to inhibit worse tendency in future in time.

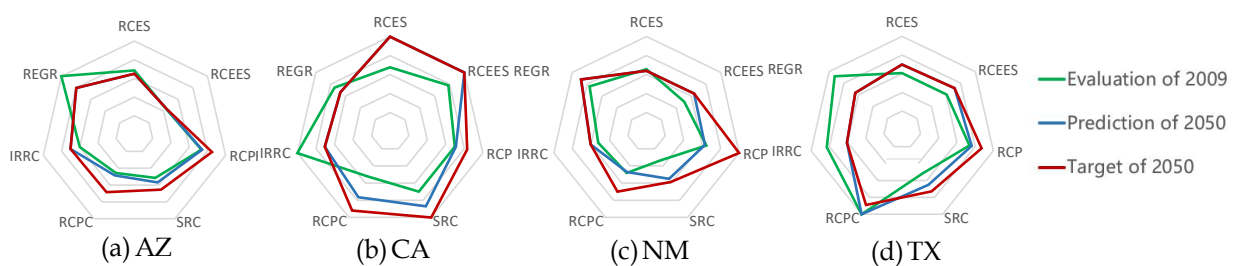Based on the above principles, we propose the following 3 actions and quantitatively interpret their promising effects.

**Action 1: Stimulating RC and RP of NM and TX in short term**

According to our prediction of 2025, we discover the following decrease in RC and RP of NM and TX, listed in Table 8. We observe that in short term, the renewable energy turns a negative trend in NM and TX, which indicates that the governors of NM and TX should stimulate the promotion of renewable energy use. Particularly NM should focus on it, since it is predicted that in 2025 the renewable energy import (RI) of NM still increases by almost $(199\_93)/93 = 114\%$ compared with 2009. This shows the possible promotion of unrenewable energy use by NM governments, on which the interstate compact should exert more notice.

Table 8: The decrease of RC and RP of NM and TX in 2025

|  | NM | TX |
|---|---|---|
| **RP** | $(35635 - 30119)/35635 = 15.5\%$ | $(356635 - 283870)/356635 = 20.4\%$ |
| **RC** | $(33785 - 26700)/33785 = 21.0\%$ | $(303697 - 245937)/303697 = 19.0\%$ |

**Action 2: Reducing Unrenewable Energy Production in AZ.**

Certainly, all states should decrease their usage of unrenewable energy including motor gasoline, natural gas and petroleum products. This problem is extremely serious in AZ, as there is a steep decline of the share of renewable production (SRP): with a value of 30% in 2025 and 87% in 2050. This sharp decrease echoes the historic evolution of unrenewable energy production of AZ, with an almost 800 times groundbreaking increase from 556 (1960) to 482422 (2009) billion Btu. Both strongly show that the wide industrial and social need of unrenewable energy in AZ. Therefore the to realize the goal of AZ, AZ should limit the production of unrenewable energy, especially natural gas and gasoline accounting for the principal portions (20.2% and 28.9% currently).

**Action 3: Promoting Diverse Usage of More Kinds of Renewable Energy in NM, TX**

We note from our prediction that NM and TX may both have a decrease in RCES, i.e. the variety of different kinds of renewable energy consumption. The predicted decrease is shown in Table 9, which may be a consistent downhill in future. To inhibit such worse tendency, NM and TX can try other possible renewable energy and expand its consumption variety. Currently solar and geothermal energy is of the lowest utilization rate in renewable energy, which account for only 1.3% and 4.4%. Therefore, motivating the use of these energy has great potentials in enhancing renewable energy usage.

In all, according to our current profile assessment and prediction, CA performs the best and can develop as its current trend without exerting more additional actions. Therefore, the compact can make CA and other three states cooperate with each other. Specifically, for example, CA can provide more renewable export to AZ to stimulate the renewable energy consumption there, and to NM and TX to impede the declining trend of their renewable consumption.

Table 9: The decrease of RCES of NM and TX in 2025 and 2050

|  | NM | TX |
|---|---|---|
| 2025 | $(0.746 - 0.55)/0.746 = 26.3\%$ | $(0.614 - 0.38)/0.614 = 38.1\%$ |
| 2050 | $(0.746 - 0.43)/0.746 = 42.3\%$ | $(0.614 - 0.25)/0.614 = 59.3\%$ |

# 6  Sensitivity Analysis

We probe into the sensitivity of some parameters in our ARMA-GPR models. As shown in Fig. 13(a), when we change $b$ from 0.1 to 0.9, the predicted curve moves upward. The prediction of 2025 varies much while the prediction of 2050 remains unchanged. This indicates that the short-term prediction is sensitive to $b$ and the long-term prediction is insensitive. The reason is that the ARMA prediction counts little in long-term prediction. Moreover, Fig. 13(b) shows the variation when we change $c_0$ from 0.1 to 0.9. As we can see, with $c_0$ increasing, the ARMA-GPR prediction curve approaches GPR prediction curve. This is because larger $c_0$ generates faster decaying of the effects by ARMA and makes the curve converge to GPR prediction more quickly.



*(a)* sensitivity of $b$                                  (b) sensitivity of $c_j$
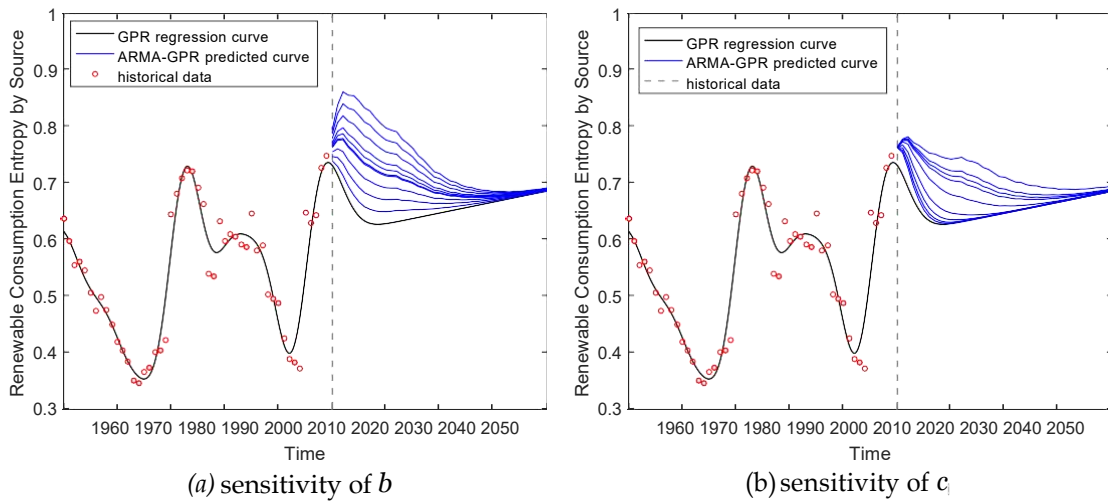
Figure 13: Sensitivity Analysis of Parameters in ARMA-GPR model

Then we probe into the sensitivity of our optimization model in Section 5.1. Setting different , we obtain the optimized value of the objective function, shown in Table 10. The

Table 10: Sensitivity Analysis of Parameter in Optimization Model

| $\lambda$ | 0.011 | 0.012 | 0.013 | 0.014 | 0.015 | 0.016 | 0.017 | 0.018 | 0.019 |
|---|---|---|---|---|---|---|---|---|---|
| $TRCP^j$ | 3.11 | 4.25 | 6.45 | 7.43 | 9.23 | 11.25 | 14.50 | 16.34 | 18.33 |

results in the table show that different  bring about much variation of the optimized value. In fact,  represents the weight of difference between total renewable consumption and production. If this weight is small, then the target RC given by the optimization will increase much. This provides more flexibility to the governors. Specifically, if the governors desire to set a high and challenging target,  can be set as a relatively small value. To achieve this

target, each state must take more actions and enhance cooperation. Hence, one can set a proper to adjust the difficulty of the compact target.

# 7    Strengths and Weaknesses

## 7.1    Strengths

- **Low Feature and Evaluation Complexity:** We refine 20 out of total 605 features, and further generalize 7 criteria integrated into 1 in evaluation;

- **High Generalizability:** We apply GPR to characterize all items in energy profiles with different trends due to its non-parametic property;

- **Bilateral Similarity Analysis:** We seminally dissect similarity of energy profiles among different states by the combination of GRA and KRCC, with concentration on value and tendency similarity respectively;

- **Temporally Adaptive Prediction:** We are avant-couriers proposing the ARMA-GPR Hybrid model suitable in both short-term and long-term prediction, by adaptively altering the weights of local and global information;

- **Quantified and Rational Goals:** We set quantified goals strictly based on optimization theory.

## 7.2    Weaknesses

- **No Verification of Raw Data:** We have no guarantee of the accuracy of given data [5].

- **No Involvement of Other States:** We do not consider states other than CA, AZ, NM, and TX, due to the lack of relevant data.

- **No Consideration of State Self-Interest:** We set our targets based on maximizing total renewable energy consumption, but in fact each state may care more about its own development while regardless of others, thus not following our mutual targets.

# 8    Conclusion

In this paper, we first create a energy profile by selecting and aggregating the variables in the provided data. Then we propose a framework called CAFE (Characterization, Analysis, Forecasting, and Evaluation of Energy Profile) to characterize the evolution of energy profile time series, analyze the similarity of energy profile among four states, evaluate the renewable energy usage, and forecast the future evolution of energy profile. Based on the results of CAFE, we adopt multi-objective programming to set up renewable energy usage targets for CA, AZ, NMïijŇand TX. Moreover, we propose a series of concrete actions for each state to meet the targets. Finally, we conduct sensitivity analysis of some parameters in our model and discuss the strengths and weakness of our work.

# MEMORANDUM

**To:** Governors of CA, AZ, NM, and TX

**From:** Team #72969

**Date:** Feb 12$^{th}$, 2018

**Subject:** Energy Profiles Characterization, Prediction and Future Goals

Honorable Governors of CA, AZ, NM, and TX,

Currently the four states intend to enact an energy compact, with emphasis on policies of renewable energy usage. As the person in charge, our team has made a comprehensive study on renewable energy use based on historical data from 1960 to 2009, including characterizing the energy profile of four states in this period, predicting energy use to 2050, and setting future goals and actions.

**Energy Profiles:** In 2009, both the renewable energy production and consumption in all states increase at least 100% and 200% compared with 1960 respectively. Main difference exists in unrenewable energy production: CA and TX decreases by 16.5% and 3.4%, while NM and AZ increases by 52% and 800 times. It renders CA and TX an increasing share of renewable energy production, while AZ a decrease by 85%. Generally, CA is judged to have the best energy profile in 2009.

**Predictions:** We make predictions of energy profiles in 2025 and 2050. For the renewable energy production and consumption, both NM and TX first decrease before 2025 while increase around 2050, while CA and AZ consistently increase. Other factors are of reasonable future trends expect: (i) the share of renewable energy production, where AZ is predicted to have a decrease of 30%/87% in 2025/2050 compared with that in 2009; (ii) the variety of renewable energy consumption, where NM and TX continues to decrease by 37%/49% and 38%/59%.

**Goals:** Renewable energy consumption (RC) takes the dominant role of all renewable energy use, acting as the linchpin of renewable energy promotion. Therefore the states should stimulate RC in future. Based on our prediction, the goals of RC are respectively CA: 988787/1227217, AZ: 148028/175154, NM: 30120/36359, TX: 283871/371206 (billion Btw) in the year of 2025/2050.

**Actions:** Our main proposed actions, ranked by significance, are: (i) Stimulating RC and RP of NM and TX in short term; (ii) Reducing Unrenewable Energy Production in AZ; (iii) Promoting Diverse Usage of More Kinds of Renewable Energy in NM and TX. All are based on profiles as of 2009 and our predictions. Meanwhile, CA can perform positively without any extra action, so CA can export more renewable energy to the other three states to stimulate their RCs.

The above is the summary of our study. We sincerely hope that it will provide you with useful information.

Thanks!

# References

[1] J. F. Zimmerman, *Interstate cooperation: Compacts and administrative agreements*. SUNY Press, 2012.

[2] "Explanation of energy profiles." http://libguides.gatech.edu/c.php?g=54047p=349537.

[3] A. Cammarano, C. Petrioli, and D. Spenza, "Pro-energy: A novel energy prediction model for solar and wind energy-harvesting wireless sensor networks," in *IEEE 9th International Conference on Mobile Adhoc and Sensor Systems (MASS)*, pp. 75–83, 2012.

[4] J. Virote and R. Neves-Silva, "Stochastic models for building energy prediction based on occupant behavior assessment," *Energy and Buildings*, vol. 53, pp. 183–193, 2012.

[5] "Dataset source." https://catalog.data.gov/dataset/state-energy-data-system-seds-complete-dataset-through-2009sec-dates.

[6] U. Çaydaş and A. Hasçalık, "Use of the grey relational analysis to determine optimum laser cutting parameters with multi-performance characteristics," *Optics & Laser Technology*, vol. 40, no. 7, pp. 987–994, 2008.

[7] Y. Kuo, T. Yang, and G.-W. Huang, "The use of grey relational analysis in solving multiple attribute decision-making problems," *Computers & industrial engineering*, vol. 55, no. 1, pp. 80–93, 2008.

[8] "Grey relational analysis." https://en.wikipedia.org/wiki/Grey-relational-analysis.

[9] S. Akhtar and S. Rozi, "An autoregressive integrated moving average model for short-term prediction of hepatitis c virus seropositivity among male volunteer blood donors in karachi, pakistan," *World Journal of Gastroenterology (WJG)*, vol. 15, no. 13, pp. 1607–1612, 2009.

## Appendix A: Similarity calculation MATLAB code

```matlab
factor = [423970 295243 315198 696241;
4.70 0.35 0.20 2.50;
880 1250 1740 520 ;
563 323 350 734 ;
17 24 15 .5 21;
92.6 22 6.62 40.6;
2424 295 91 1648 ;
61925 43269 43551 59995];

F = 8 ;
factorrelation = zeros(N,N,F);

factort = [];
similarity1t = [];
for f = 1 :F

figure ;
x = [ ];
y = [ ];
for n1 = 1 :N
for n2 = 1 :N
if n1~=n2
factorrelation(n1,n2,f)= abs(factor(f,n1)-factor(f,n2))/factor(f,n1);
plot(factorrelation(n1,n2,f),similarity1(n1,n2),'bo');
x = [x factorrelation(n1,n2,f)];
y = [y similarity1(n1,n2)];
hold on ;
end
end
end
factort = [factort;x];
similarity1t = [similarity1t;y];
[w,s]= polyfit(x,y,1);
x = 0:0.01:100;
y = w(1)*x+w(2);
plot(x,y);
a=0;b1 = 1.2*max(max(factorrelation(:,:,f)));b2 = 1.2*max(max(similarity1));
%xlim([a b1]);
%ylim([0.8*min(min(similarity1)) b2]);
axis([a b1 a b2]);
end

for f = 1 :F
tmp = factorrelation(:,:,f);
factorkendallc(f)= corr(similarity2(:),tmp(:),'type','pearson');
```

```
46   end
47   tmp1 = [ ] ;
48   f o r f = 1 : F
49   tmp = [ ] ;
50   f o r n1 = 1 :N
51   tmp = [ tmp ; f a c t o r r e l a t i o n ( n1 , : , f ) ' ] ;
52   end
53   tmp1 = [ tmp1 tmp ] ;
54   end
55   f a c t o r k e n d a l l p = p a r t i a l c o r r ( [ tmp1 , s i m i l a r i t y 2 ( : ) ] ) ;
```

## Appendix B: GPR regression MATLAB code

```
1    N = 4 ;
2    K =  20 ;
3    T =  50 ;
4
5    weight  = z e r o s (N,K) ;
6    bi as = z e r o s (N,K) ;
7    Sigma =   z e r o s  (N,K) ;
8    deviati on = z e r o s (N,K) ;
9
10   x = 1960 : 2009 ;
11   x = x ' ;
12   te s tx  = 2010 :2050 ;
13   te s tx = testx ' ;
14   f o r k=1:20
15
16   d a t a f i t = datanew (6*k −4:6*k −1 ,2:T+1) ;
17   d a t a f i t ( isnan ( d a t a f i t ) ) = 0;
18
19   y = d a t a f i t ( 1 , : ) ;
20   gprMdl1 = f i t r g p ( x , y , ' Basis ' , ' l i n e a r ' , . . .
21   ' FitMethod ' , ' exact ' , ' PredictMethod ' , ' exact ' , ' Kernel Function ' , ' squaredexponential ' , '
         Standardize ' , 1 , ' Re gul ari zati on ' , 0.2 ) ;
22
23   y = d a t a f i t ( 2 , : ) ;
24   gprMdl2 = f i t r g p ( x , y , ' Basis ' , ' l i n e a r ' , . . .
25   ' FitMethod ' , ' exact ' , ' PredictMethod ' , ' exact ' , ' Kernel Function ' , ' squaredexponential ' , '
         Standardize ' , 1 , ' Re gul ari zati on ' , 0.2 ) ;
26   beta2 = gprMdl2 . Beta ( 1 ) ;
27
28   y = d a t a f i t ( 3 , : ) ;
29   gprMdl3 = f i t r g p ( x , y , ' Basis ' , ' l i n e a r ' , . . .
30   ' FitMethod ' , ' exact ' , ' PredictMethod ' , ' exact ' , ' Kernel Function ' , ' squaredexponential ' , '
         Standardize ' , 1 , ' Re gul ari zati on ' , 0.2 ) ;
31   beta3 = gprMdl3 . Beta ( 1 ) ;
32
```

```matlab
33  y = datafit(4,:);
34  gprMdl4 = fitrgp(x,y,'Basis','linear',...
35  'FitMethod','exact','PredictMethod','exact','KernelFunction','squaredexponential','
        Standardize',1,'Regularization',0.2);
36  beta4 = gprMdl4.Beta(1);
37
38  testx = 1960:2009;
39  [ygr1,bi1,ci1] = predict(gprMdl1,testx');
40  [ygr2,bi2,ci2] = predict(gprMdl2,testx');
41  [ygr3,bi3,ci3] = predict(gprMdl3,testx');
42  [ygr4,bi4,ci4] = predict(gprMdl4,testx');
43
44  error(k) = mean(abs(ygr1' - datafit(1,:))./(datafit(1,:)));
45
46  weight(1,k) = gprMdl1.Beta(1);
47  bias(1,k) = gprMdl1.Beta(2);
48  Sigma(1,k) = gprMdl1.Sigma;
49  deviation(1,k) = mean((bi1(1:50)))/mean(datafit(1,:));
50  weight(2,k) = gprMdl2.Beta(1);
51  bias(2,k) = gprMdl2.Beta(2);
52  Sigma(2,k) = gprMdl2.Sigma;
53  deviation(2,k) = mean((bi2(1:50)))/mean(datafit(2,:));
54  weight(3,k) = gprMdl3.Beta(1);
55  bias(3,k) = gprMdl3.Beta(2);
56  Sigma(3,k) = gprMdl3.Sigma;
57  deviation(3,k) = mean((bi3(1:50)))/mean(datafit(3,:));
58  weight(4,k) = gprMdl4.Beta(1);
59  bias(4,k) = gprMdl4.Beta(2);
60  Sigma(4,k) = gprMdl4.Sigma;
61  deviation(4,k) = mean((bi4(1:50)))/mean(datafit(4,:));
62
63  subplot(2,10,k)
64  plot(testx,ygr1,'b','LineWidth',1,'Color','r');
65  set(gca,'xtick',[],'xticklabel',[]);
66  set(gca,'ytick',[],'yticklabel',[]);
67  xlim([1960 2009]);
68  hold on
69  end
```

## Appendix C: ARMA-GPR regression MATLAB code

```matlab
1  N = 4;
2  K = 20;
3  T = 50;
4
5  x = 1960:2009;
6  x = x';
7  testx = 2010:2050;
```

```matlab
8    te s tx = testx ';
9  %f o r k=1:K;
10   k=13;
11   d a t a f i t = datanew (6*k-4:6*k-1,2:T+1);
12   d a t a f i t ( isnan ( d a t a f i t ) ) = 0;
13
14   y = d a t a f i t (1 ,:);
15   y0 = [ diff ( d a t a f i t (1 ,:) ) ';repmat(y(1,49),41,1)];
16   z = id d a t a ( y0 );
17  %z . y = cumsum( z . y ) ;%i n te g r a te d data
18   m = armax( z (1:4 9 ) , [ 3 3 ] , ' In te g rate N o i s e ' , true );
19   yma1 = p r e d i c t (m, z , 4 1 );
20   yma1 = yma1 . outputdata ( 5 0 : 9 0 );
21   gprMdl1 = f i t r g p (x,y,' Basis ',' lin e ar ' ,...
22   ' FitMethod ',' exact ',' PredictMethod ',' exact ',' Kernel Function ',' squaredexponential ','
            Standardize ' ,1,' Re gul ari zati on ' ,0.2);
23   beta1 = gprMdl1 . Beta ( 1 );
24
25   y = d a t a f i t (2 ,:);
26   y0 = [ diff ( d a t a f i t (2 ,:) ) ';repmat(y(1,49),41,1)];
27   z = id d a t a ( y0 );
28  %z . y = cumsum( z . y ) ;%i n te g r a te d data
29   m = armax( z (1:4 9 ) , [ 3 3 ] , ' In te g rate N o i s e ' , true );
30   yma2 = p r e d i c t (m, z , 4 1 );
31   yma2 = yma2 . outputdata ( 5 0 : 9 0 );
32   gprMdl2 = f i t r g p (x,y,' Basis ',' lin e ar ' ,...
33   ' FitMethod ',' exact ',' PredictMethod ',' exact ',' Kernel Function ',' squaredexponential ','
            Standardize ' ,1,' Re gul ari zati on ' ,0.2);
34   beta2 = gprMdl2 . Beta ( 1 );
35
36   y = d a t a f i t (3 ,:);
37   y0 = [ diff ( d a t a f i t (3 ,:) ) ';repmat(y(1,49),41,1)];
38   z = id d a t a ( y0 );
39  %z . y = cumsum( z . y ) ;%i n te g r a te d data
40   m = armax( z (1:4 9 ) , [ 3 3 ] , ' In te g rate N o i s e ' , true );
41   yma3 = p r e d i c t (m, z , 4 1 );
42   yma3 = yma3 . outputdata ( 5 0 : 9 0 );
43   gprMdl3 = f i t r g p (x,y,' Basis ',' lin e ar ' ,...
44   ' FitMethod ',' exact ',' PredictMethod ',' exact ',' Kernel Function ',' squaredexponential ','
            Standardize ' ,1,' Re gul ari zati on ' ,0.2);
45   beta3 = gprMdl3 . Beta ( 1 );
46
47   y = d a t a f i t (4 ,:);
48   y0 = [ diff ( d a t a f i t (4 ,:) ) ';repmat(y(1,49),41,1)];
49   z = id d a t a ( y0 );
50  %z . y = cumsum( z . y ) ;%i n te g r a te d data
51   m = armax( z (1:4 9 ) , [ 3 3 ] , ' In te g rate N o i s e ' , true );
```

```matlab
52  yma4 =  predict(m, z, 41);
53  yma4 = yma4.outputdata(50:90);
54  gprMdl4 = fitrgp(x,y,'Basis','linear',...
55  'FitMethod','exact','PredictMethod','exact','Kernel Function','squaredexponential','...
        Standardize',1,'Regularization',0.2);
56  beta4 = gprMdl4.Beta(1);
57
58  alpha = 0.8;
59  ysp1 = alpha*cumsum(yma1) + (1-alpha)*(prelation(1,2,k)*cumsum(yma2)+prelation(1,3,
        k)*cumsum(yma3)+prelation(1,4,k)*cumsum(yma4))/sqrt(prelation(1,2,k)^2+prelation
        (1,3,k)^2+prelation(1,4,k)^2);
60  ysp2 = alpha*cumsum(yma2) + (1-alpha)*(prelation(2,1,k)*cumsum(yma1)+prelation(2,3,
        k)*cumsum(yma3)+prelation(2,4,k)*cumsum(yma4))/sqrt(prelation(2,1,k)^2+prelation
        (2,3,k)^2+prelation(2,4,k)^2);
61  ysp3 = alpha*cumsum(yma3) + (1-alpha)*(prelation(3,1,k)*cumsum(yma1)+prelation(3,2,
        k)*cumsum(yma2)+prelation(3,4,k)*cumsum(yma4))/sqrt(prelation(3,1,k)^2+prelation
        (3,2,k)^2+prelation(3,4,k)^2);
62  ysp4 = alpha*cumsum(yma4) + (1-alpha)*(prelation(4,1,k)*cumsum(yma1)+prelation(4,2,
        k)*cumsum(yma2)+prelation(4,3,k)*cumsum(yma3))/sqrt(prelation(4,1,k)^2+prelation
        (4,2,k)^2+prelation(4,3,k)^2);
63
64  [ygr1,bi1,ci1] = predict(gprMdl1,testx);
65  [ygr2,bi2,ci2] = predict(gprMdl2,testx);
66  [ygr3,bi3,ci3] = predict(gprMdl3,testx);
67  [ygr4,bi4,ci4] = predict(gprMdl4,testx);
68
69  a1 = 0.4; a2 = 0.5; decay = -0.6;
70  %ypred1 = ((1 - a1.* exp(decay.*(testx -2010)))).* ygr1 + a1.* exp(decay.*(testx -2010)).*
        ysp1);
71  ypred1 = (1-a1.* exp(decay.*(testx -2010))).* ygr1 + a1.* exp(decay.*(testx -2010)).*
        (datafit(1,50)+ysp1);
72  ypred2 = (1-a1.* exp(decay.*(testx -2010))).* ygr2 + a1.* exp(decay.*(testx -2010)).*
        (datafit(1,50)+ysp2);
73  ypred3 = (1-a1.* exp(decay.*(testx -2010))).* ygr3 + a1.* exp(decay.*(testx -2010)).*
        (datafit(1,50)+ysp3);
74  ypred4 = (1-a1.* exp(decay.*(testx -2010))).* ygr4 + a1.* exp(decay.*(testx -2010)).*
        (datafit(1,50)+ysp4);
75  end
```