2013

Analysis Of Sensor Data In Cyber-physical System

Xianglong Kong
University of Massachusetts Amherst

ANALYSIS OF SENSOR DATA IN CYBER-PHYSICAL SYSTEM

A Thesis Presented

by

XIANGLONG KONG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL AND COMPUTER ENGINEERING

September 2013

Electrical and Computer Engineering

ANALYSIS OF SENSOR DATA IN CYBER-PHYSICAL SYSTEM

A Thesis Presented

by

XIANGLONG KONG

Approved as to style and content by:

Tilman Wolf, Chair

Weibo Gong, Member

Michael Zink, Member

C.V.Hollot, Department Head
Electrical and Computer Engineering

ACKNOWLEDGEMENT

First and foremost, I would like to express the deepest appreciation to my advisor, Professor Tilman Wolf, who has the attitude and the substance of a genius and who inspired my interest in computer networking. Without his guidance and persistent help this thesis would not have been possible. He always encourages me and inspires me with lots of valuable and insights ideas.

I am heartily thankful to Professor Weibo Gong, and Professor Michael Zink, for their constructive advice and invaluable help on both of my research and future career.

I also would like to acknowledge the members in the Network Systems Lab for the knowledge and experience they have shared with me. In particular, I want to thank Nauman Javed and Cong Wang, who worked together with me on this great project, and helped me a lot.

Finally, I appreciate all of the sincere support from my family and my friends, this thesis pales in comparison to what I gained from them.

ABSTRACT

ANALYSIS OF SENSOR DATA IN CYBER-PHYSICAL SYSTEM

SEPTEMBER 2013

XIANGLONG KONG

B.E, DALIAN MARITIME UNIVERSITY

M.S.E.C.E, UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Tilman Wolf

Cyber-Physical System (CPS) becomes more and more importance from industrial application (e.g., aircraft control, automation management) to societal challenges (e.g. health caring, environment monitoring). It has traditionally been designed to one specific application domain and to be managed by a single entity, implemented communication between physical world and computational world. However, it still just work within its domain, and not be interoperability. How to make it into scalable? How to make it reusing? These questions become more and more necessary. In this paper, we are trying to developing a common CPS infrastructure, let it be an innovative CPS crossing multiple domains to broad use sensors and actuators. Here, we implement a technique for automatically build a model according to the sensor data in different domains. And based on our approach under continuous situation, it could identify the sensor values right now or estimate next few time step, which we call spatial model or temporal model.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 The Evolution of Cyber-Physical System

In general, CPS contains three parts, sensor, computation application, and actuator. It performs three basic functions: sensing characteristics of the physical world, computing appropriately based on sensor input and other data sources, and generating response actions through actuation. The basic principles of operation are illustrated in Figure 1.1.

With the rapid development, Cyber-Physical Systems (CPS) is distributing to many industries and business companies. CPS represent the technical foundation to solve some important social and environmental problems. These domain range from focused control problems (e.g. industrial automation, aircraft control, etc.) to larger-scale problems (e.g. environmental monitoring, health care, etc.).

Traditionally, the CPS has been designed to one specific application domain and to be managed by a single entity. In addition, its components are deployed and managed by

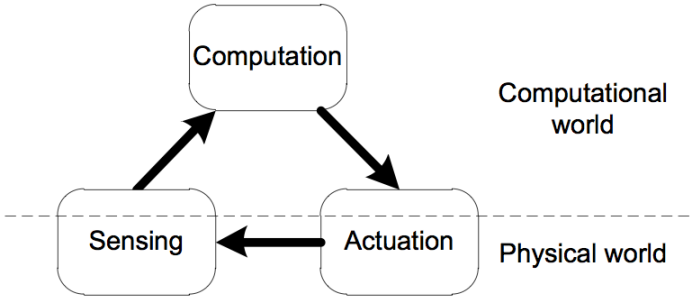

Figure 1.1: Principles of operation of cyber-physical systems

the same administrative entity. It is dominated by vertically integrated "stovepipe" design. Obviously, as indicated in the name cyber-physical, there are two key components of cyber-physical system are "information"(cyber) and "computational part"(modeling the physics of the system).[11]

But the problem is to set up a scalable deployment of cyber-physical system since the cost of sensors and actuators becomes prohibitively expensive. So developing a common CPS infrastructure becomes crucial for the innovative CPS to solve scalability problem. Imaging that, if we could use sensor information and actuators from multiple domains across domains, the cost of the devices interfacing with the physical world will be amortized by many applications, then the scalability problem could be solved. We could achieve horizontal integration such that sensing information could be shared across different applications, multiple computation services could coexist, and the response could be controlled by different entities. In computation world, computational components are captured from "Internet of Things" (IoT), which draws on an analogy many different computers interact with common data communication network.

Due to Internet of Things managed by different administrative entities, it raises many new interesting and technical problems:

Interoperability: CPS components need to exchange data and operate together to form a CPS solution.

Security and Trust across domains: It is important to establish trust between different entities because they are controlled by separate entities. Here we introduce a concept called Verification layer, which is used to implement techniques to check if sensors report correct data and actuators implement correct responses. In addition, establishing security is also a non-ignorable problem here, it could share allow for verification between components under different administrative control.

Economics: It couples with the identification of any associated synergies, utilize to model horizontal integration, and incentivize the entities for sharing sensor information and access to actuators.

In my approach, I address one specific problem that build verification between entities: how to detect outliers in the CPS sensor information. We need build an outlier detection system that could evaluate sensor readings to decide if this data can be trusted or to determine which ones are outliers in the context of other sensor readings. In our outlier detection system, the outliers could show up as an extreme values (correct sensing information), or a malicious intent (incorrect sensing information) according to the computational components. As we known, the sensors obtain information from physical world that are usually continuous and look smoothing or similar values over a short time, such as air temperature, barometric pressure, traffic congestion, and so on. Our detection makes full use of multiple sources of sensor information to determine reported observations match the temporal and spatial context. Those incorrect sensing information will be deleted as untrusted sensors. But those correct sensing information process to report on property in physical world because it shows the values change drastically over a short time slot, appearing some irregular phenomenon.

To implement our outlier detection in large-scale Internet of Things and make it broadly applicable to many different areas, we need to analyze different entities correlationship, select the principal entities from all sensor information that we could use. After determining which types of sensor information to use for outlier detection, we are trying to find what kind of model to use to present the physical phenomenon.

To prove our approach effectiveness, we test it in one of domains called weather sensing. We collect different types of sensor information, such as air temperature, dew temperature, barometric pressure, and visibility. Thus, we select different principal components according to different regions we plan to observe, and develop internal model of weather appropriately, then determine which readings are considered as outlier in our outlier detection system.

1.2 Objectives

In this proposal, we analyze the cross-correlation of all types sensor sources and select some of them as the principal components to build our model. The first challenge is to discover these relationships. Since the significance of different sensor readings, their polynomial degrees, and their interaction, toward determining a specific dependent sensor reading are not known. We use a statistical method called "step-wise regression" to search for the appropriate regression model. We evaluate and provide a preferable internal model that could represent the physical phenomenon.

In the next step of this thesis, we try to use outlier detection system to evaluate sensor readings within a temporal and spatial context. Then, we evaluate and demonstrate the effectiveness of the proposed technique in a weather sensing scenario.

1.3 Contributions

Our main contributions are as below:

1. Analyze different types of sensor sources and determine to selection of each variable.

2. Evaluate and compare the performance of different internal models with multiple regression method and holt-winters smoothing method to determine one could represent the physical phenomenon.

3. Design of an outlier detection system that evaluates sensor readings within a temporal and spatial context.

1.4 Thesis Organization

The rest of the thesis is organized as follows: Chapter 2 presents some background and related work on cyber-physical system. Chapter 3 introduces the method of data analysis, preprocessing, and data cleaning. Then, we figure out one most efficient spatial model

through consider all possible situations. The evaluation of outlier detection system based on different methods will be presented at Chapter 4. Chapter 5 implements an approach that could predict next few hour time step sensor value. Evaluation of our model technique is done in Chapter 6.

CHAPTER 2

BACKGROUND

In this chapter, we will provide a detailed description and review for some of the previous research works on cyber-physical system.

2.1 Traditional Cyber-Physical System

In this section, we will describe recent CPS researches in conference publications. Many existing CPS solutions are based on stovepipe architectures. These studies include many applications in road traffic management[4], energy system[8] [6], power infrastructure[1], and health care(monitoring devices[2]).

It basically focus on cyberizing the physical and physicalizing the cyber. The challenge of integrating computing and physical processes has been recognized for some time in cyberizing part. Time synchronization is different implemented in software execution since there is no statement like "current time is t" and no semantic notion of time passing. In physicalizing part, what CPS needs is not faster computing, but physical actions taken at the right time[7]. It prefer to building a reliable system with strong robustness and adaptation, not only faster computing.

Many researches focus on the development of underlying technologies such as embedded system design, new software or hardware verification and system design. For example, transportation systems could benefit considerably from better embedded intelligence in automobiles, components and runtime substrates also have been discussed in control computing, and real-time environments. And the security(e.g. robustness, data reliability, and

fault-tolerance) in CPS, have been explored in various contexts(e.g. control system, system resilience). But most of them are limited in their own specific domains.

Detecting anomalies method has also been derived in other domains. For example, it used to detect network anomalies by comparing the current network traffic against a baseline distribution in network traffic[12]. We propose this technique to combine anomaly detection information from multiple sources to get more accurate results. In our approach, we identify anomalies

2.2 Distributed Sensing in the Internet of Things

In many cases, the physical environment need to sense by a group of distributed sensors, and each of them senses the data independently. All of them change continuously with time, and could spread to all geographic pictures. For example, weather stations deploy over a certain geographical area and sense individually one or more weather variables (e.g. air temperature, pressure, visibility, and dew temperature etc.) in weather sensing.

Assuming that there are n sensors distributed in a geographical space, which record n different sensor information in the same variable of interest. And there are k different variables in this geographical space need observation, denoted by x_1, x_2, ... , x_k. $x_(ji)$(t), denotes that sensor i senses and reports a time series of the variable x_j. For each time slot, the value of every variable x_i could represent by the set x_i(t):

$$x_i(t) = x_{i1}(t), x_{i2}(t), ..., x_{in}(t) \tag{2.1}$$

So all variables could draw its own geographical pictures at one time slot and provide coherent pictures during a continuous time period. One of important things that we need to prerequisite is to verify each value in the set.

Our main objective is to automatically check whether a specific sensor m report us an outlier according to the other sensors' report. The key to detect the outlier is according to other different variables and find their regularity. It means that, the sensor recording the variable x_i at time slot t, may be related to another sensor recording the variable x_j at time slot t, or multiple different variables, as well as to the spatial parameters. But it is difficult to find their relationships, since they might do not have any relationship at all or complicating relationship among them. In our discussion, we assume there is no prior domain-specific knowledge so that we could apply our method into more wide fields. If their regularity could be discovered from sensor values, then erroneous sensor reports could possibly be detected.

To validate our results, we apply our method on a weather sensing application. It is a distributed sensor network and contains multiple sensed variables (e.g. air temperature, dew temperature, pressure and visibility).

2.3 Temporal extrapolation

In the weather sensing forecasting, there are several complex forecasting method. A feature based forecast model using Neural Network is proposed with high degree of accuracy and could be suitably adapted for making forecasts over larger geographical areas. It builds a fixed model with five features to predict the maximum temperature and minimum temperature.[10] Another model "Numerical Weather Prediction and hybrid ARMA/ANN" uses a technique forecast from the ALADIN NWP model combined to an Auto-Regressive and Moving Average (ARMA) model.[3] A model named "Weather Forecasting System using concept of Soft Computing" constructs an image, which represents the actual data.[9] Although these models could provide excellent capability based on raw data, all of them are just focus on their specific domain, not related to other domain. In our work , we use Holt-

Winter forecasting for temporal extrapolation, which also has been used in network volumes forecasting.

CHAPTER 3

DATA ANALYSIS AND EVALUATION

3.1 Data collection

In this study we utilized a data set from government website www.weather.gov. They provide reliable sensor information from every weather station (e.g. air temperature, dew temperature, visibility, pressure, longitude, latitude and elevation), which spread over a geographical space. We apply our methodology on these multiple variables of interest to verify the effectiveness of our approach. We collect 98 weather stations deployed at different locations in New England area (contains the states of Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, and Connecticut) and New York State. It contains weather variable values reported at one hour between May 30, 2011 and June 15, 2011.

3.2 Pre-Processing

Before build the spatial model, it is crucial to preprocess the raw sensor readings in proper method. We replace missing data by their interpolation value since the sensor readings are continuous and have some regularity to it. Then, we adjust each sensor reading into a specified standard form. It is quite useful to make unites of variables comparable from zero scaling to one while measured on different scales and to equalize the relative importance of variables. And before we do the standardization, we need to adjust our sensor readings (e.g.

influential observations due to high leverage) approaching to standard normal distribution with power/log transformation.

$$Y = Y_a + (Y_b - Y_a)\frac{x - x_a}{x_b - x_a} \tag{3.1}$$

To find the regularity of different variables, we derive a concept called multiple linear regression (MLR), which is expressed in the form as follow:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d + \epsilon \tag{3.2}$$

This is a modeling technique for analyzing the relationship between a continuous (real-valued) response variable y and one or more explanatory variables x_1, x_2, ... x_n and identifying a function could estimate the conditional expectation of the response variable given the explanatory variables. The closer the R^2 statistic to the value 1, the better the estimated regression function fits the data, since a goodness of fit measurement is represented by the R^2 statistic.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{t=1}^{k}(y_i(t) - \hat{y}_i(t))^2}{\sum_{t=1}^{k}(y_i(t) - \overline{y}_i(t))^2} \tag{3.3}$$

Where SSE is the sum of squares of the residuals, SST is the sum of squares of the total (variance of the observed values), $\overline{y}_i(t)$ is the average value of $y_i(t)$.

3.3 Variable Selection

The above-mentioned observations must be carefully screened before building the relationship model. For multiple sensor readings affecting a certain variable of interest, effective variable selection is necessary among those multiple sensor readings in order to reducing the redundant sensor readings and saving time for sensor readings without any domain specific

knowledge. Assuming that there is no prior relationship among the variables of interest, the first challenge is to filter the useless variables. Here we introduce F-statistic and ANOVA to test each one variable has a statistical significant effect on specific dependent variable. Meanwhile, we check their correlation between different sensor data.

Here we provide an example of variable selection of air temperature between different variables (e.g. pressure, dew temperature, visibility, latitude, longitude and elevation). Figure 3.1 is the correlation between different sensor data readings at 16 hour, June 1, and it shows whether and how strongly pairs of variables are related, in this way, we could get their potential modeling approach. We could see that, the correlation between air temperature and pressure is relatively obvious. Figure 3.2 and figure 3.3 show the P-value and F-value, which illustrate the effectiveness and the significance of each variable in the basic linear model. It becomes evident that, although the significant effect on air temperature between other variables have changed in different time, all these variables are non-ignorable during the observation period.

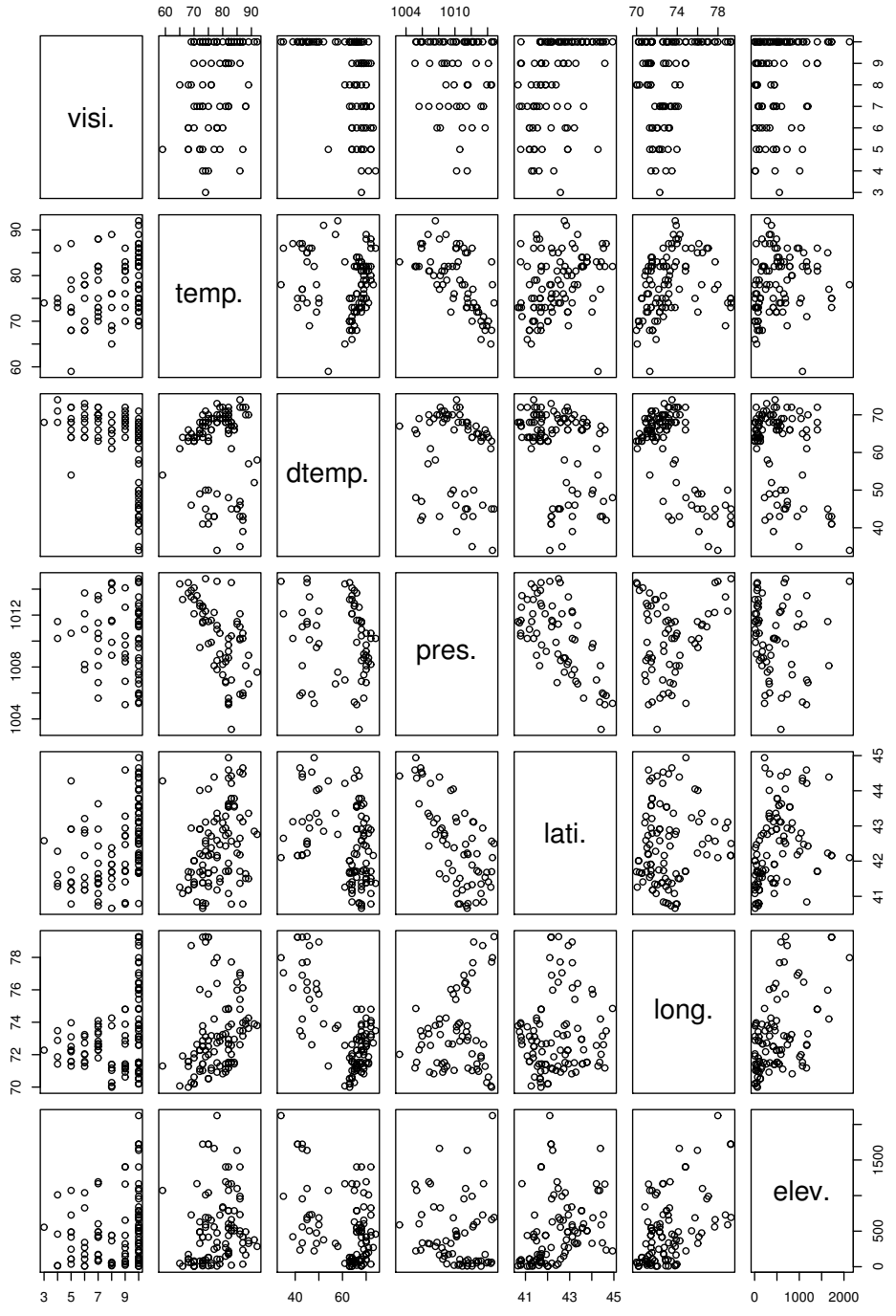

Figure 3.1: Interaction between Different Viriables

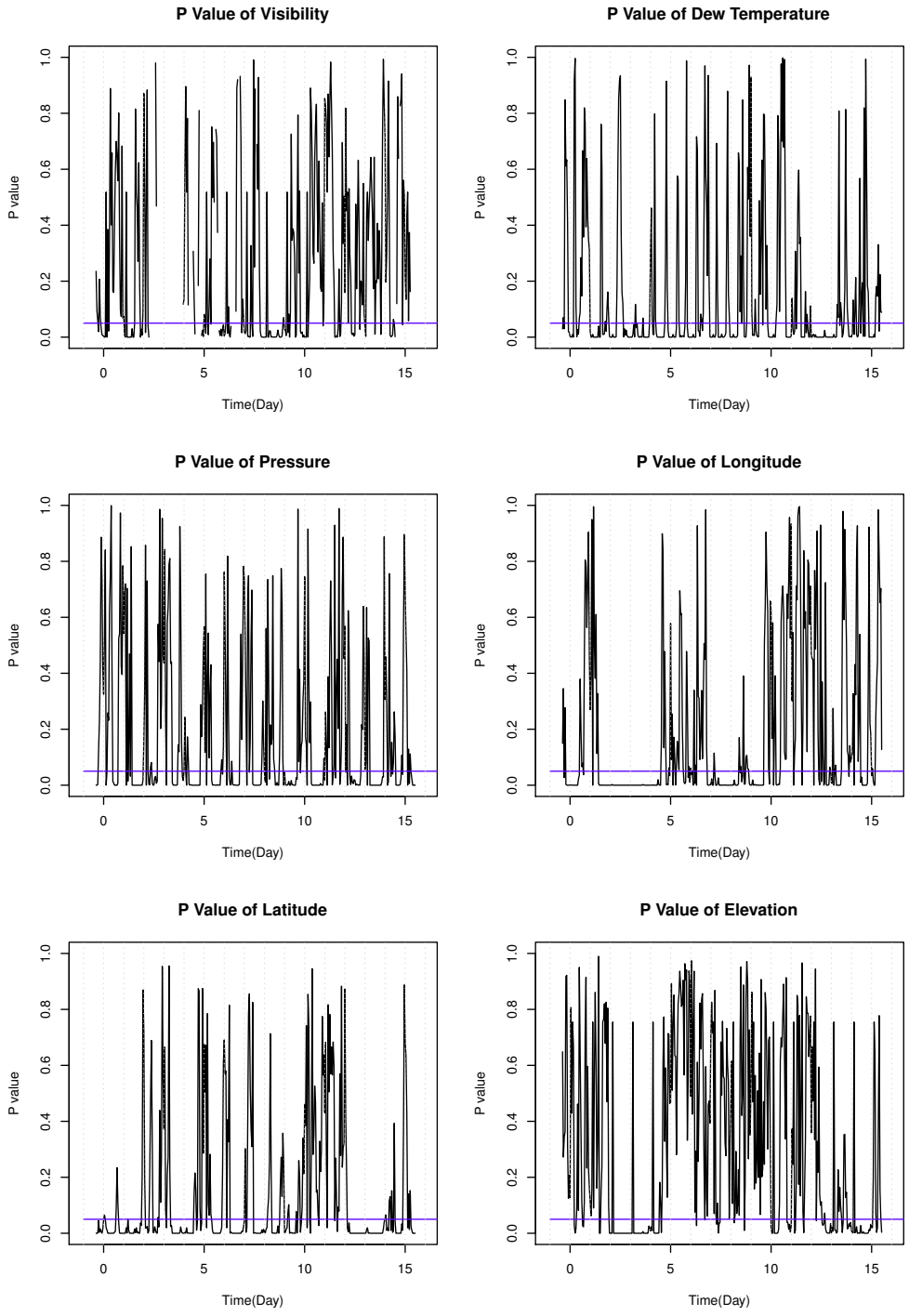

Figure 3.2: P Value of each viriable in Linear Model

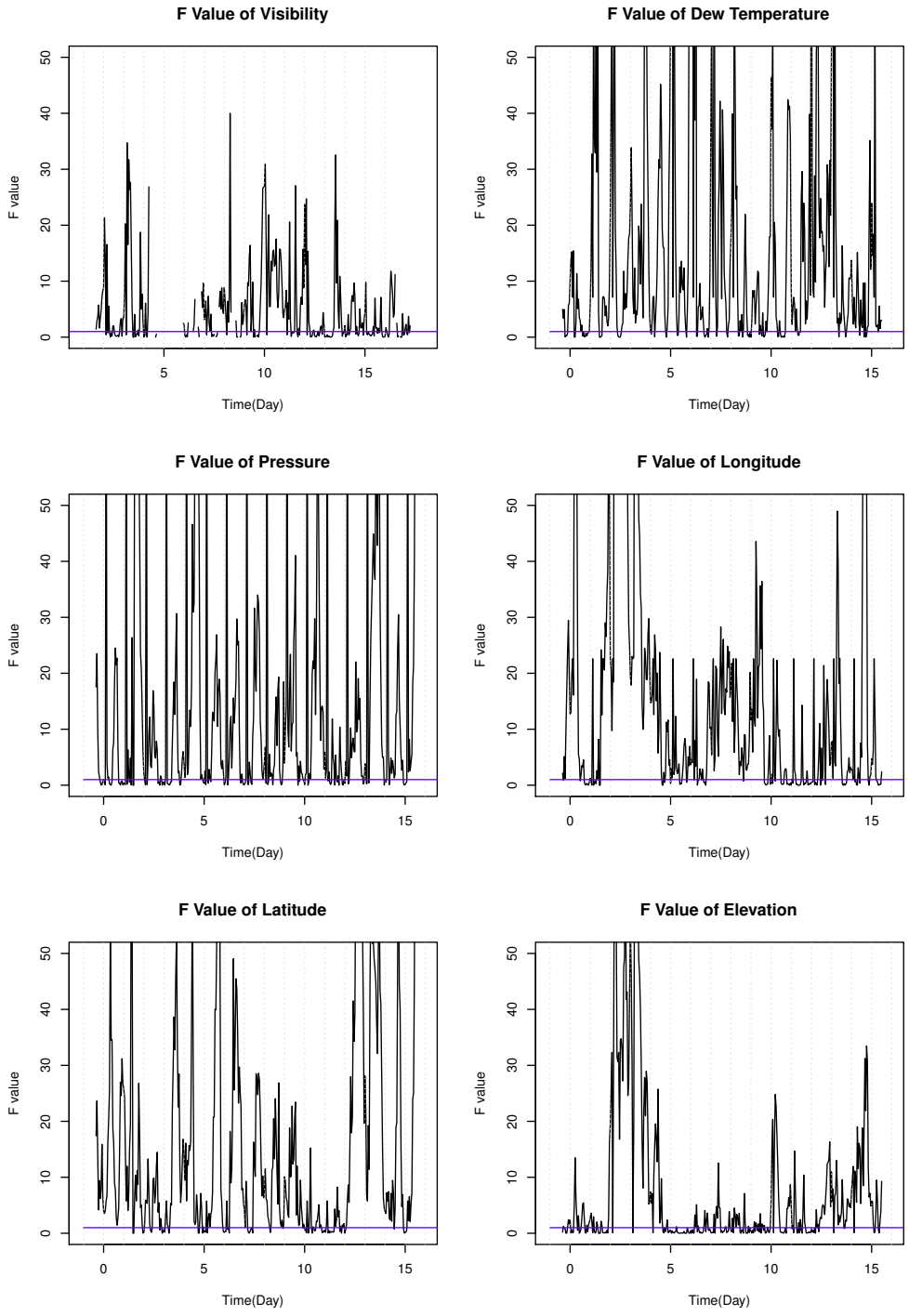

Figure 3.3: F Value of each viriable in Linear Model

CHAPTER 4

SPATIAL MODEL OF SENSOR DATA

After fitting lots of different statistical variable selection models to a given data set, we are now focus on building the model. This section details the prerequisites that how to find best fitted method after data analysis. We usually increase the degree of the polynomial till the added term is not statistically significant. At the same time, increase the interaction part also could enhance the model and decrease the error of estimated value. But here we should notify that the model with interaction term does not remove the x_1x_2 interaction term without simultaneously considering the removal of x_1^2 and x_2^2 terms, and the model with polynomial term does not eliminate lower order terms from the model even if they are not statistically significant.

We use statistical method called Akaike Information Criterion (AIC) to search for more appropriate regression model, which is a kind of "Step-wise Regression". The benefit of AIC is that it not only rewards goodness of fit (R^2 value), but includes a penalty that is an increasing function of the number of estimated parameters at the same time. This penalty discourages over fitting, regardless of the number of free parameters in data generating process.[5] It starts with all explanatory variables in the model, then remove the predictor with highest p-value greater than α_{crit}.

$$mse(\hat{y}_i) = var(y_i) + [E(\hat{y}_i) - E(y_i)]^2 \tag{4.1}$$

where $[E(\hat{y}_i) - E(y_i)]$ is called the bias in the predicting the observation y_i using $\hat{y}_i$.

$$AIC = -2ln(L) + 2k = nln(RSS/n) + 2k \tag{4.2}$$

where k is the number of parameters, and L is the likelihood function.

As an example, Figure 4.1 and figure 4.2 plot the results of models token all explanatory variables into account with different polynomial degrees, which is the estimated expression of air temperature among those sensor readings across the spatial domain of interest. We could see that, the higher degree of them do not always show as the same important as lower degree. The low degree of polynomial has the significant improvement while adding its higher degree of polynomial, but the high degree of polynomial has less significant improvement while adding its higher degree of polynomial. That means, although the estimation of air temperature do improve while the increasing the polynomial degree of the predictor variables, the improvement becomes weaken gradually while adding degree of polynomials.

Then, we use the percentage increment of AIC with different polynomial degree to evaluate the improvement capability of our model(If it less than 5%, we ignore the adding degree.). Trying to improve the model fit by doing the AIC method with different polynomials, observed which explanatory variable drop off. The figure 4.3 and figure 4.4 are evident that the percentage increment of AIC locate in 5% while polynomial degree increasing.

To compensate for improving our model, we quantify through measures of the adjusted R^2. We use adjusted R^2 to evaluate the fitness of our model. Figure 4.5 and figure 4.6 show the fitness of our model with different degrees and the difference between their adjacent degree.

4.1 Validation of Spatial Models

Then we check whether our predicted model could provide effective prediction of sensor readings from different domains. We split our observed data into the training set and test-

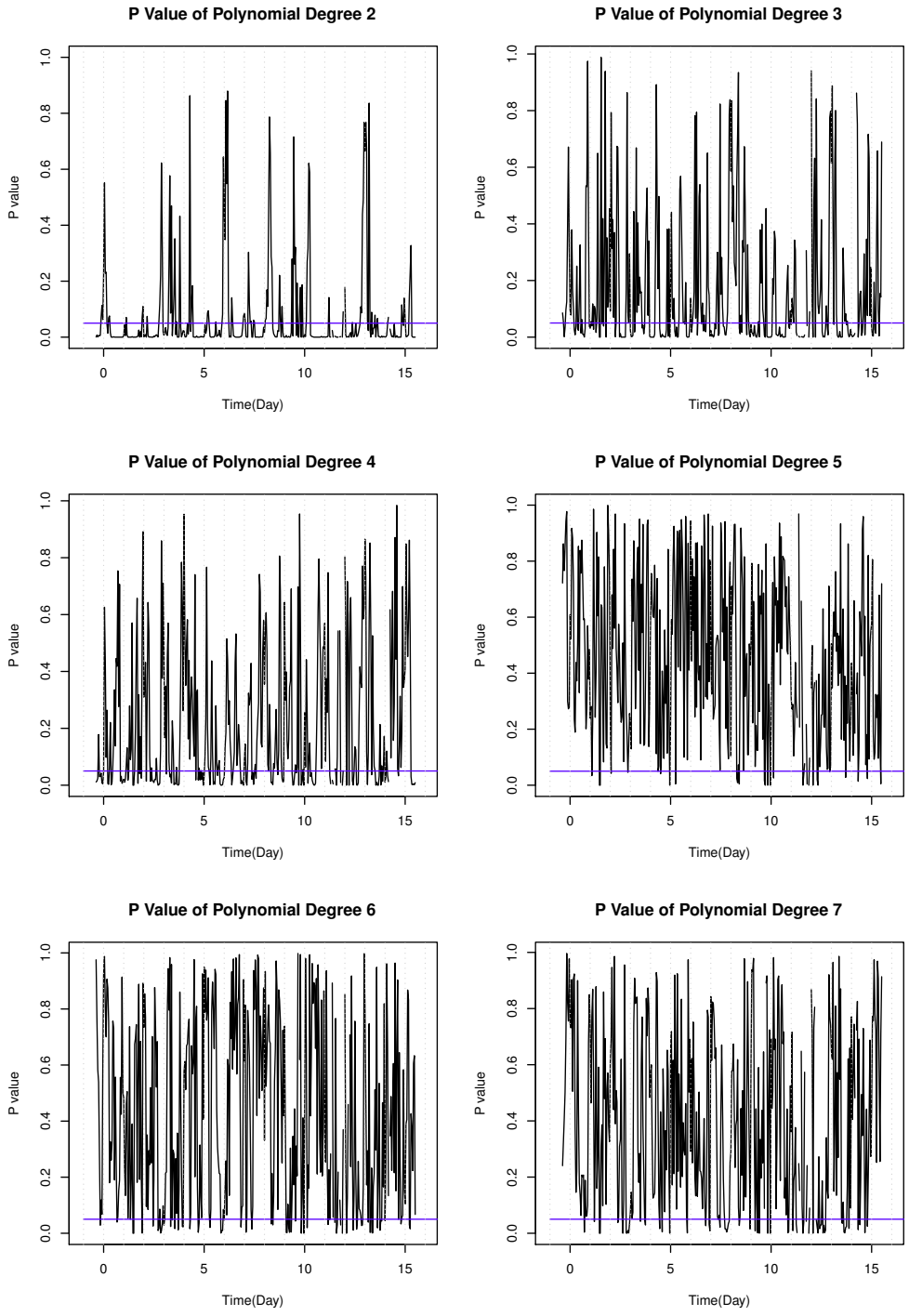

Figure 4.1: P Value with different polynomial(1)

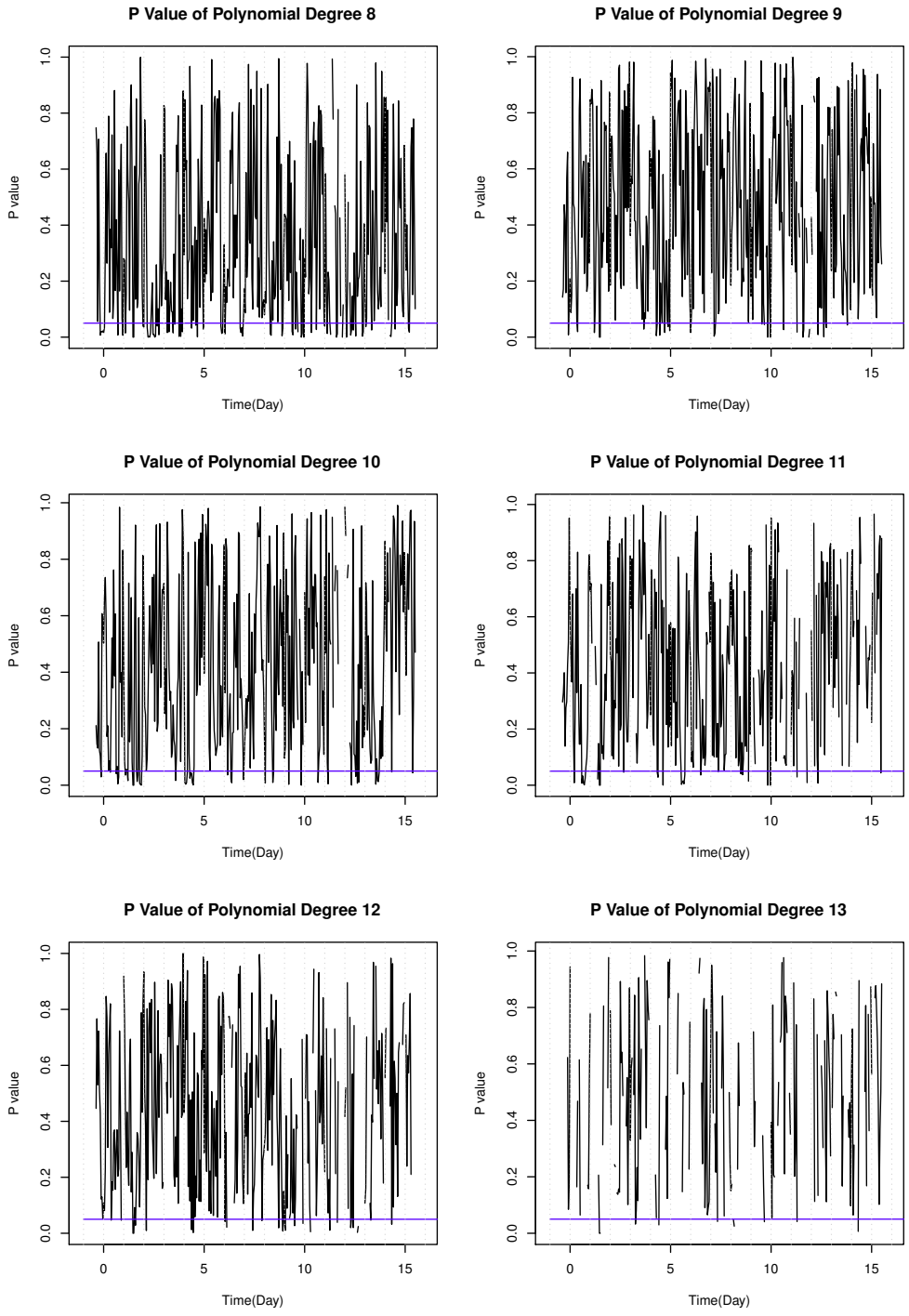

Figure 4.2: P Value with different polynomial(2)

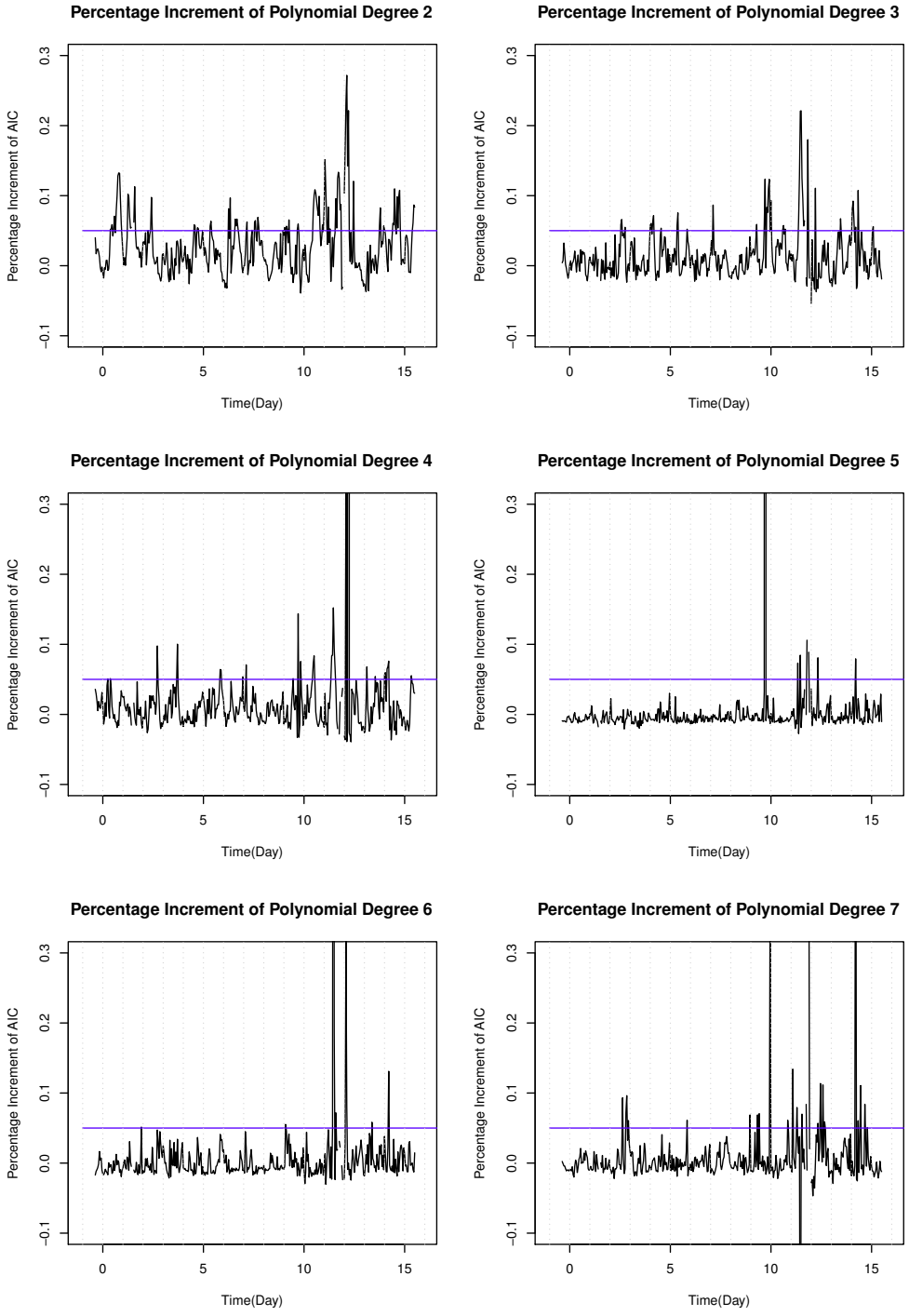

Figure 4.3: Percentage Increment of AIC with different degree polynomials(1)

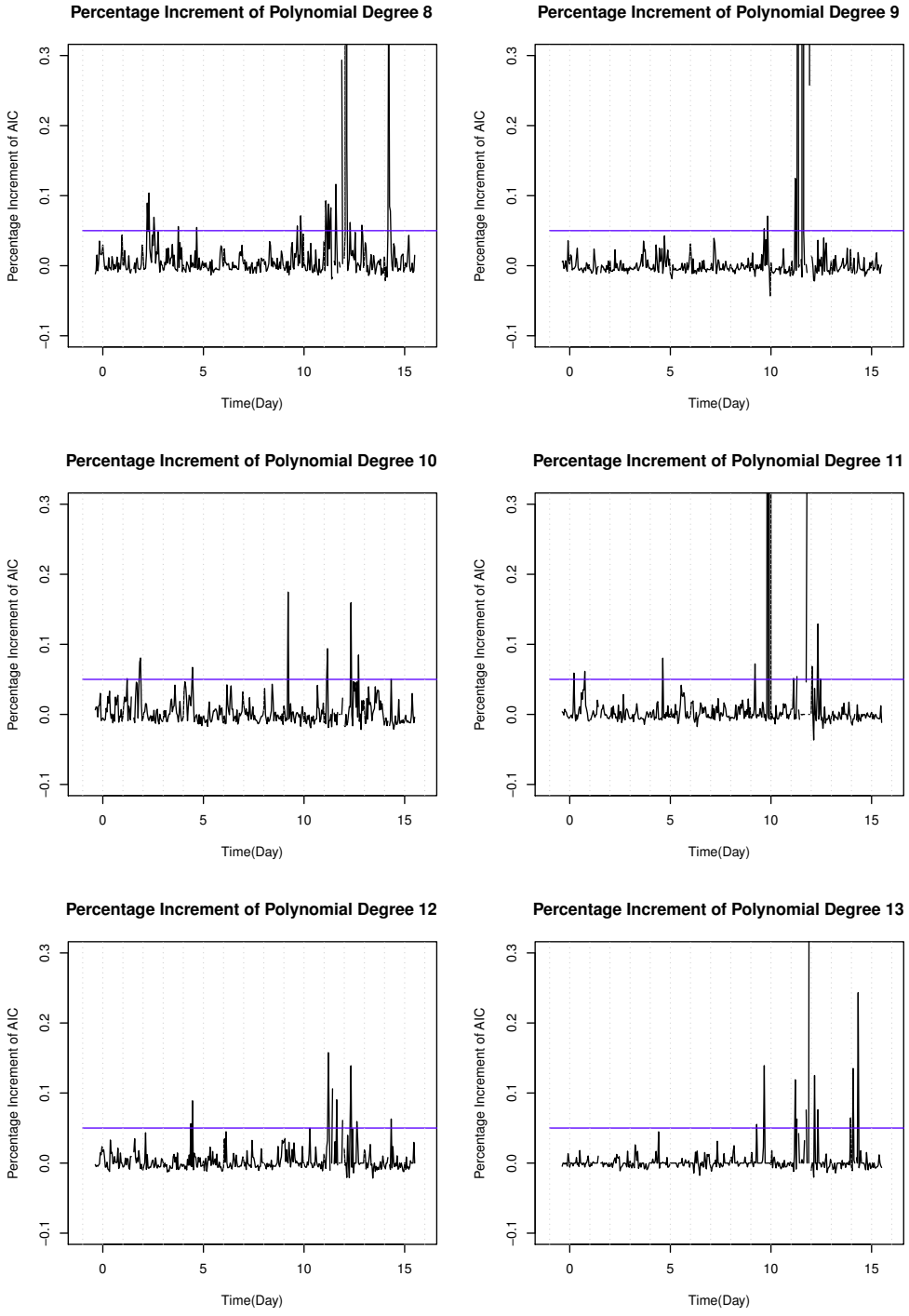

Figure 4.4: Percentage Increment of AIC with different degree polynomials(2)

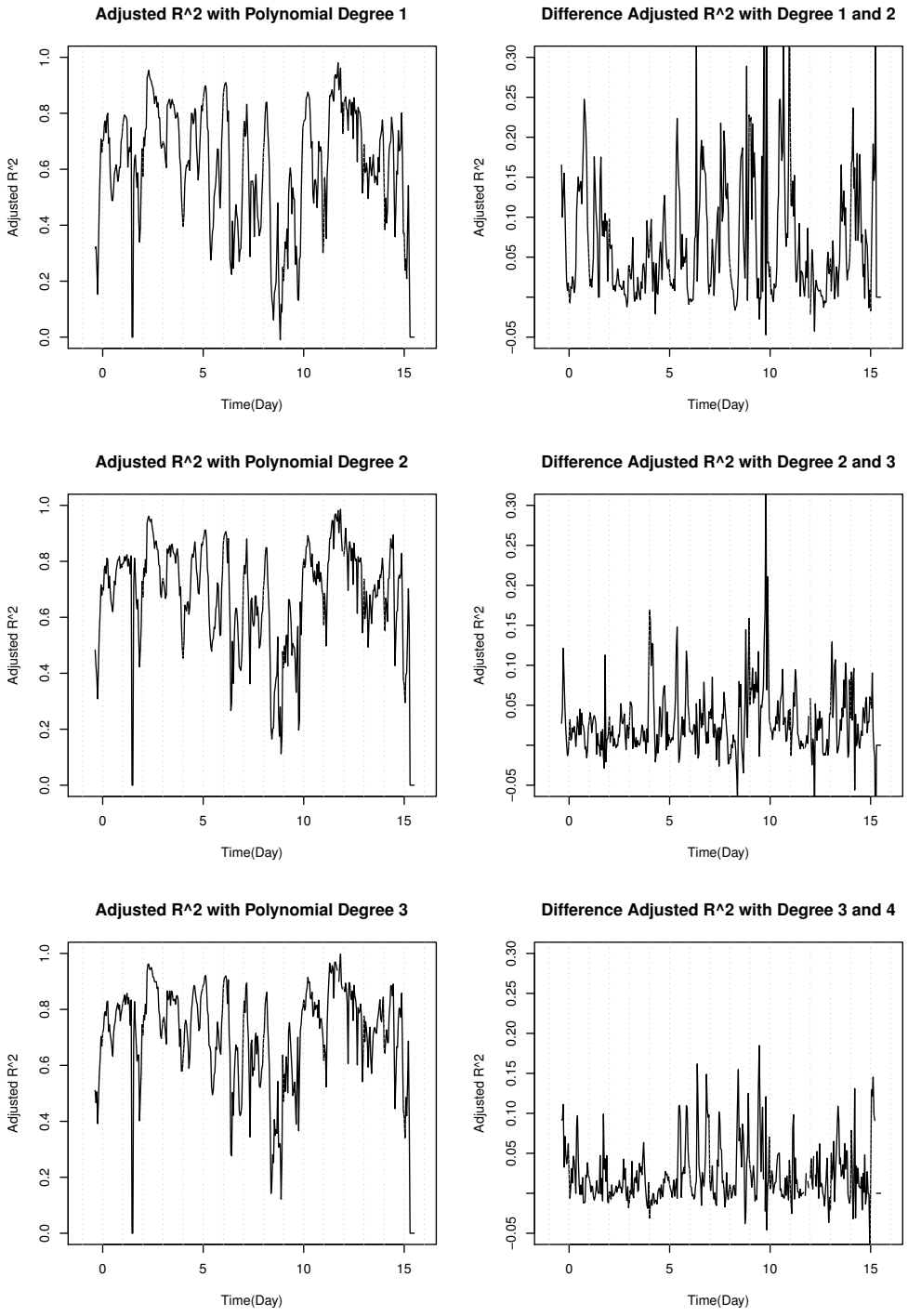

Figure 4.5: Adjusted R with Different Polynomial Degree(1)

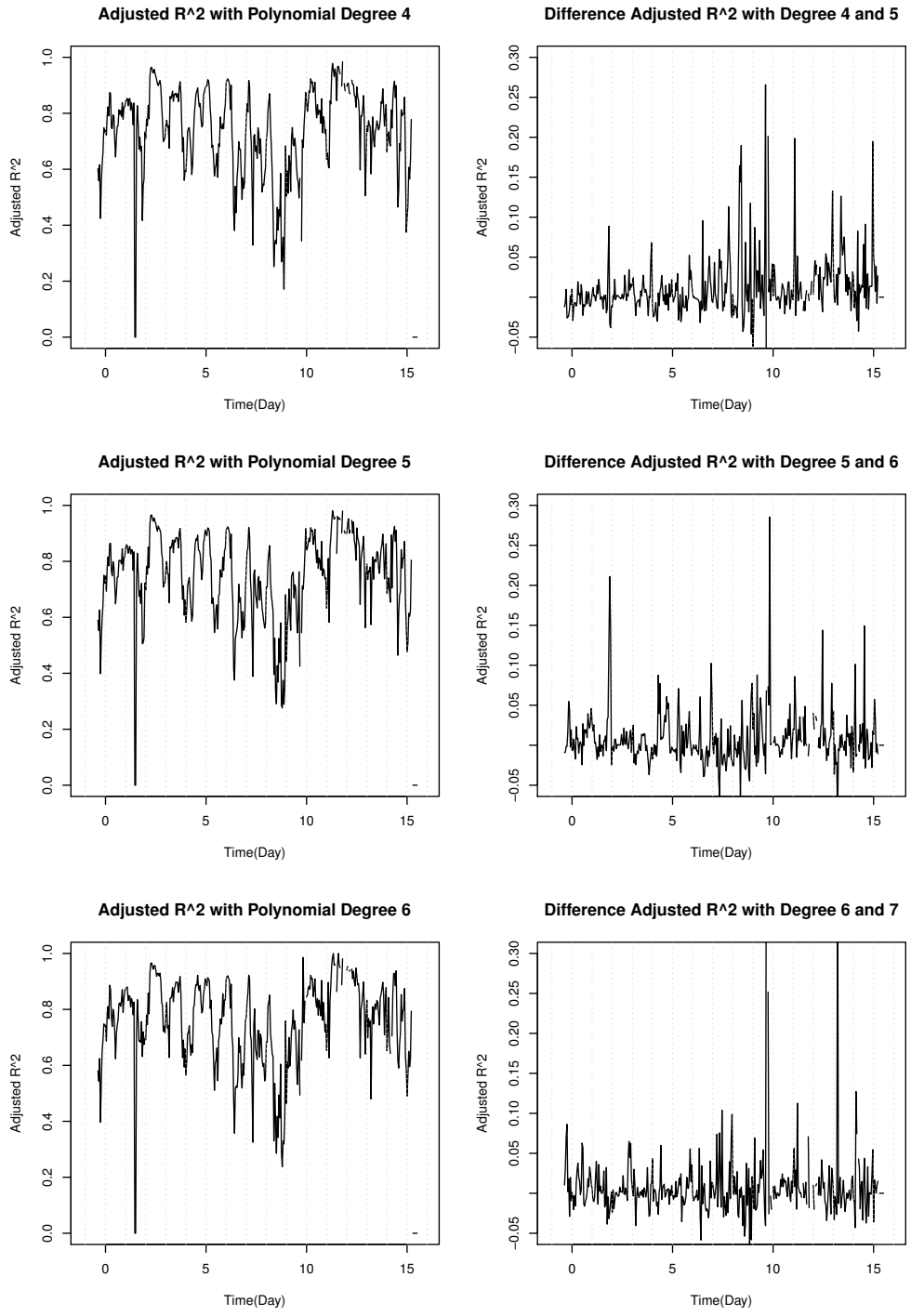

Figure 4.6: Adjusted R with Different Polynomial Degree(2)

ing set for the purpose of checking the standard error of our appropriate regression model. Although providing higher performance of adjusted R^2, the higher polynomial also leads to the larger standard error. The model is calculated using the training set, and is tested by the testing set.

Here, figure 4.7 shows the results about standard error of training set and testing set with different polynomials to estimate temperature values. Although the standard error for training set with higher polynomials becomes smaller, the standard error for testing set gets larger in some degree. That is what we do not expect. So the best model should consider the trade-off between the polynomials and standard error. To solve this problem, we need to set a threshold for the purpose of making our results more accurate and limiting increasing the degree of polynomials. Limitation of residual range is the best way.

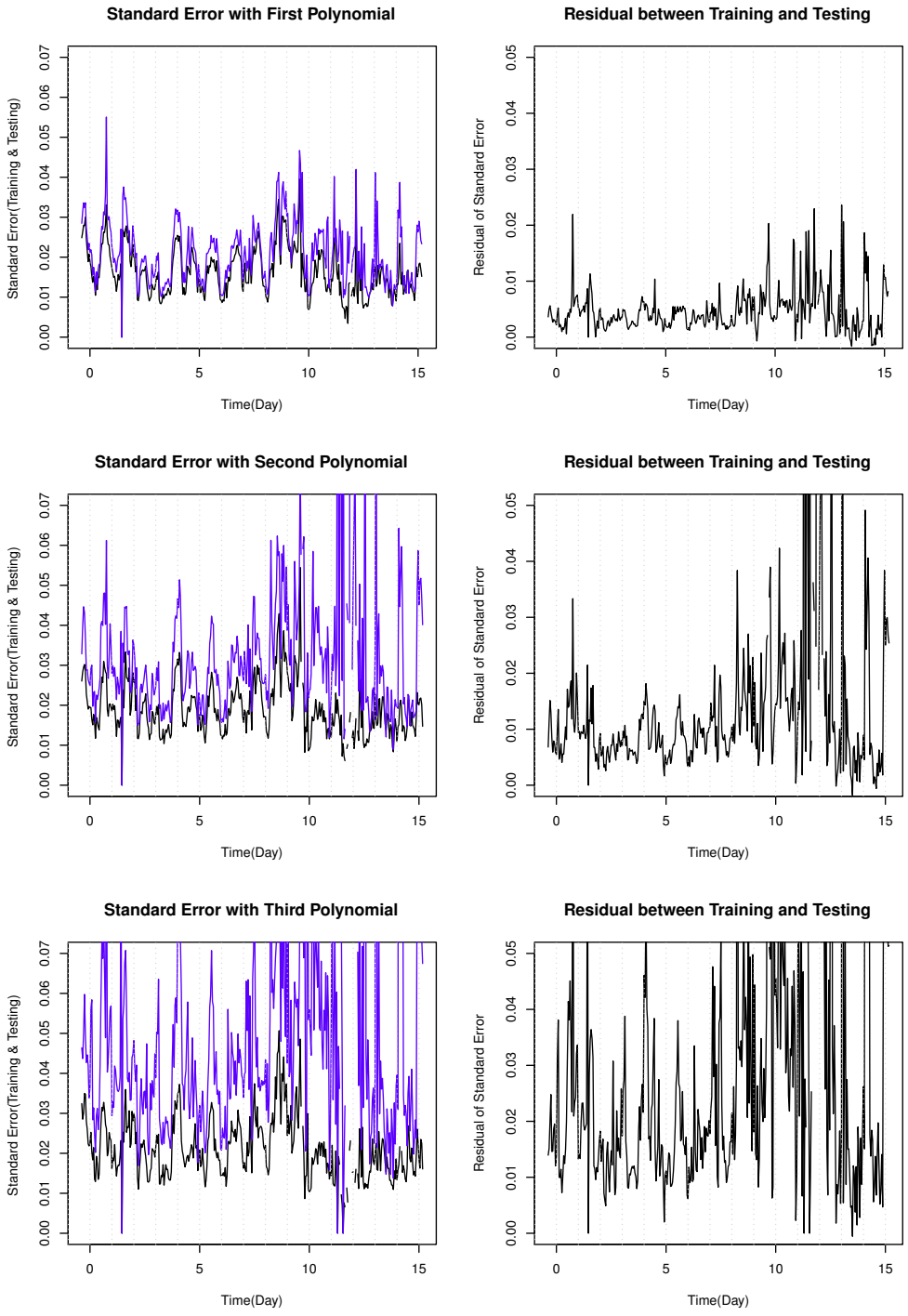

Figure 4.7: Standard Error for Different Polynomials

CHAPTER 5

TEMPORAL MODEL OF SENSOR DATA

5.1 Temporal Model with Holt-Winters smoothing

Since each correlation between those variables should be stable in some degree, we are trying to figure out their regularity from the correlations between each of them that we calculated by our outlier detection system. If we could discover the regularity of correlation between two variables in a fixed time period, then we could easily determine the expected values over time at each spatial coordinate. We use extrapolate temporally to predict sensor values for next several time steps. Here, we derive a mathematical method called Holt-Winters, which is also used for anomaly detection information to estimate network traffic volumes, since it could show the regularity of these correlationship among different sensor readings. The following equations is used when the data exhibits additive seasonality, which means the tendency of data to exhibit behavior that repeats itself every m period, and also exhibits trend, which is a smoothed estimate of average growth at the end of each period.

$$\hat{y}_{t+h|t} = l_t + b_t h + s_{t+h-m} \tag{5.1}$$

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{5.2}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{5.3}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \tag{5.4}$$

Where, I_t is the base signal also called the permanent component, b_t is a linear trend component, s_{t+h-m} is an additive seasonal factor.

And the following plots will show more details about how Holt-Winters Exponential Smoothing works in prediction. These three plots represent alpha, beta and gamma values with different coefficients in the first degree of polynomial model. α is the smoothing factor, also is called simple weighted average of the previous observation x_{t-1} and the previous smoothed statistic s_{t-1}. When the α closes to 1, it means the output has greater weight to recent changes in the data. It shows that the variable visibility has a greater smoothing effect and is less responsive to recent changes. β is the trend smoothing factor, used to estimate the difference between two successive estimates of the deseasonalized level. Here, all the variables coefficients is without trend. γ is the seasonal change smoothing factor, used to estimate the difference between two estimates in successive periods. In our plot, just visibility variable shows less effect on seasonal trend.

Thus, our model is basically implemented to interpolate spatially or to extrapolate temporally. The interpolation spatially is to verify the correctness of the other sources, and the extrapolation temporally is to predict sensor values for next several time steps.

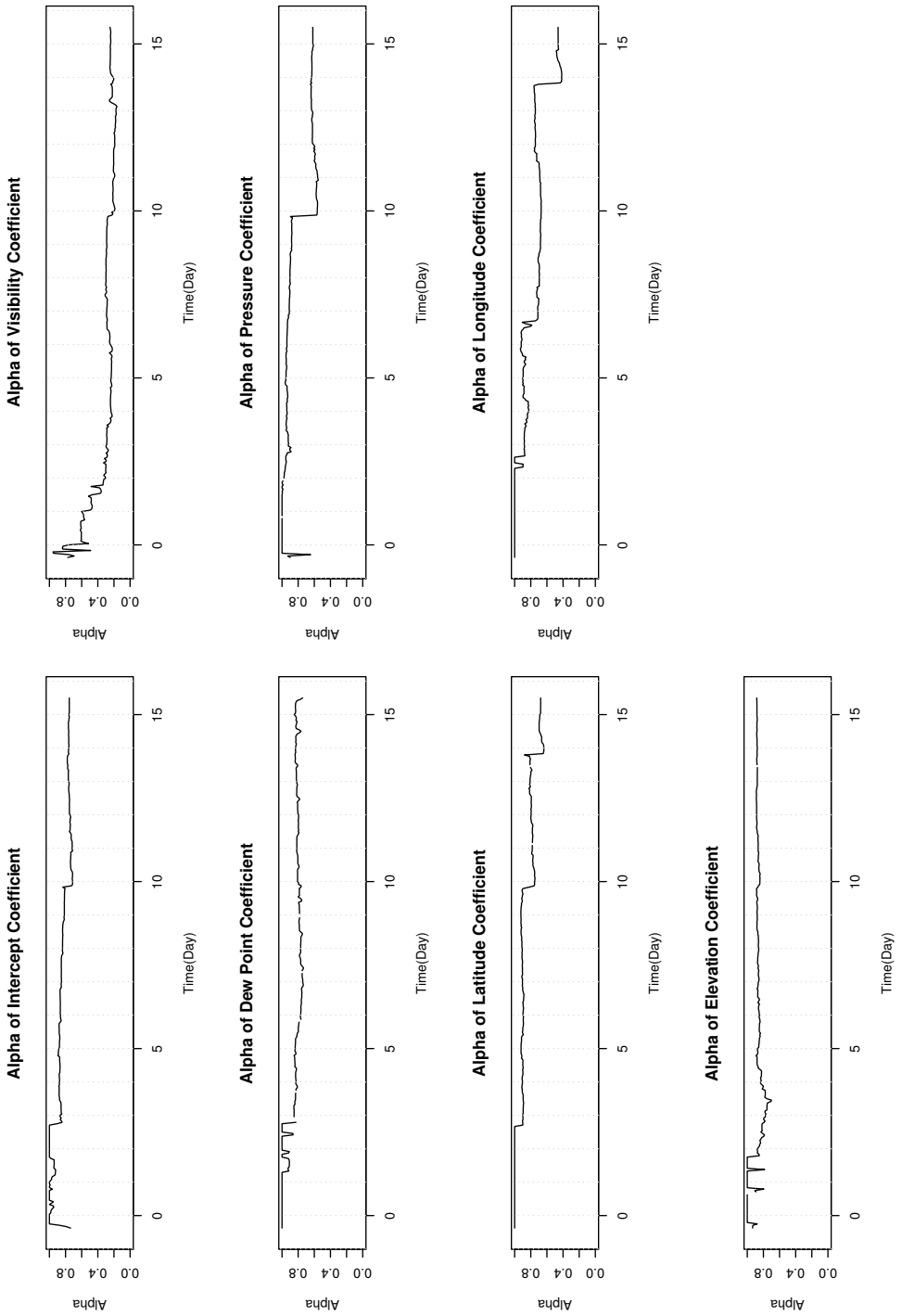

Figure 5.1: Different Coefficients Alpha with First Order Polynomial

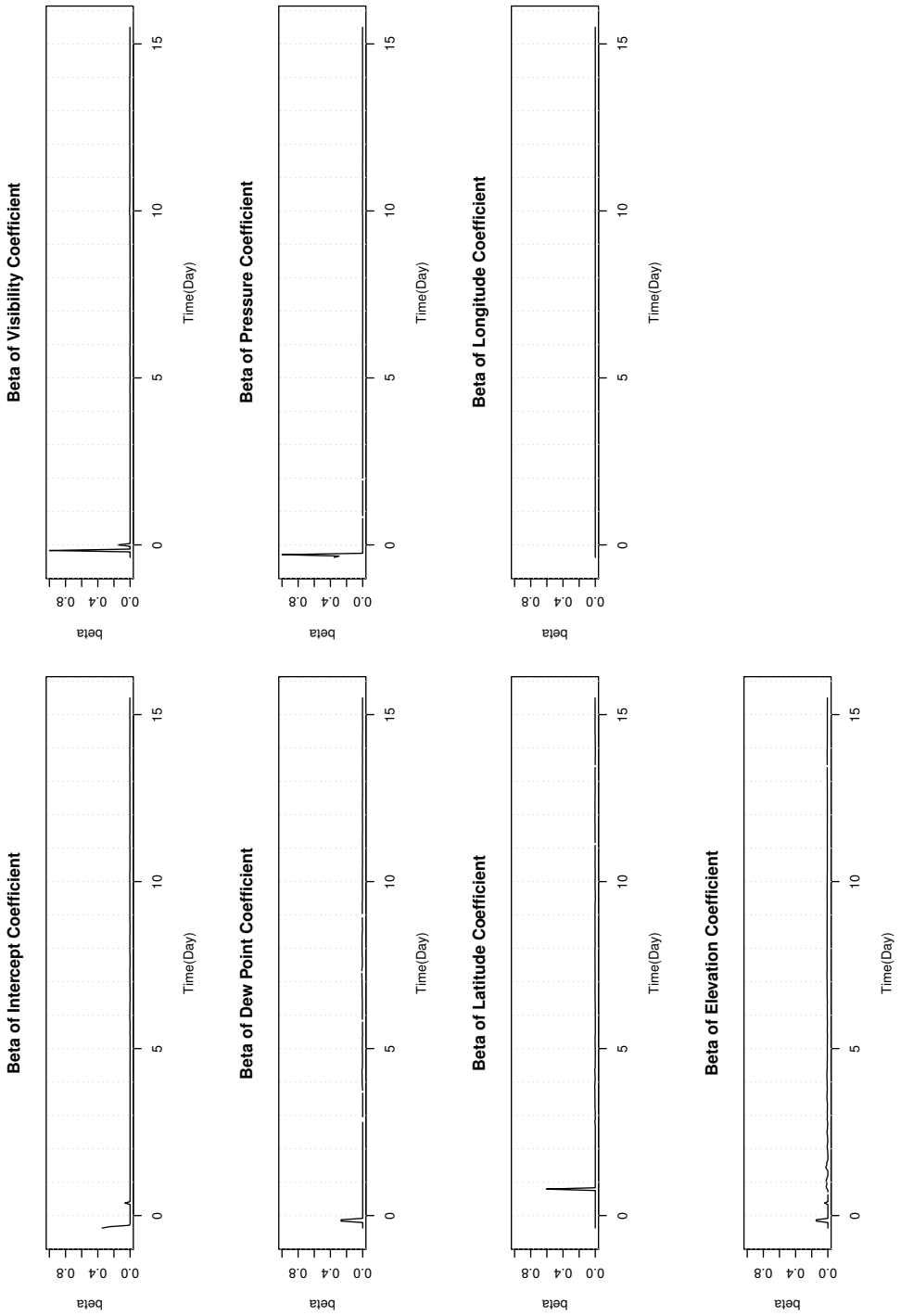

Figure 5.2: Different Coefficients Beta with First Order Polynomial

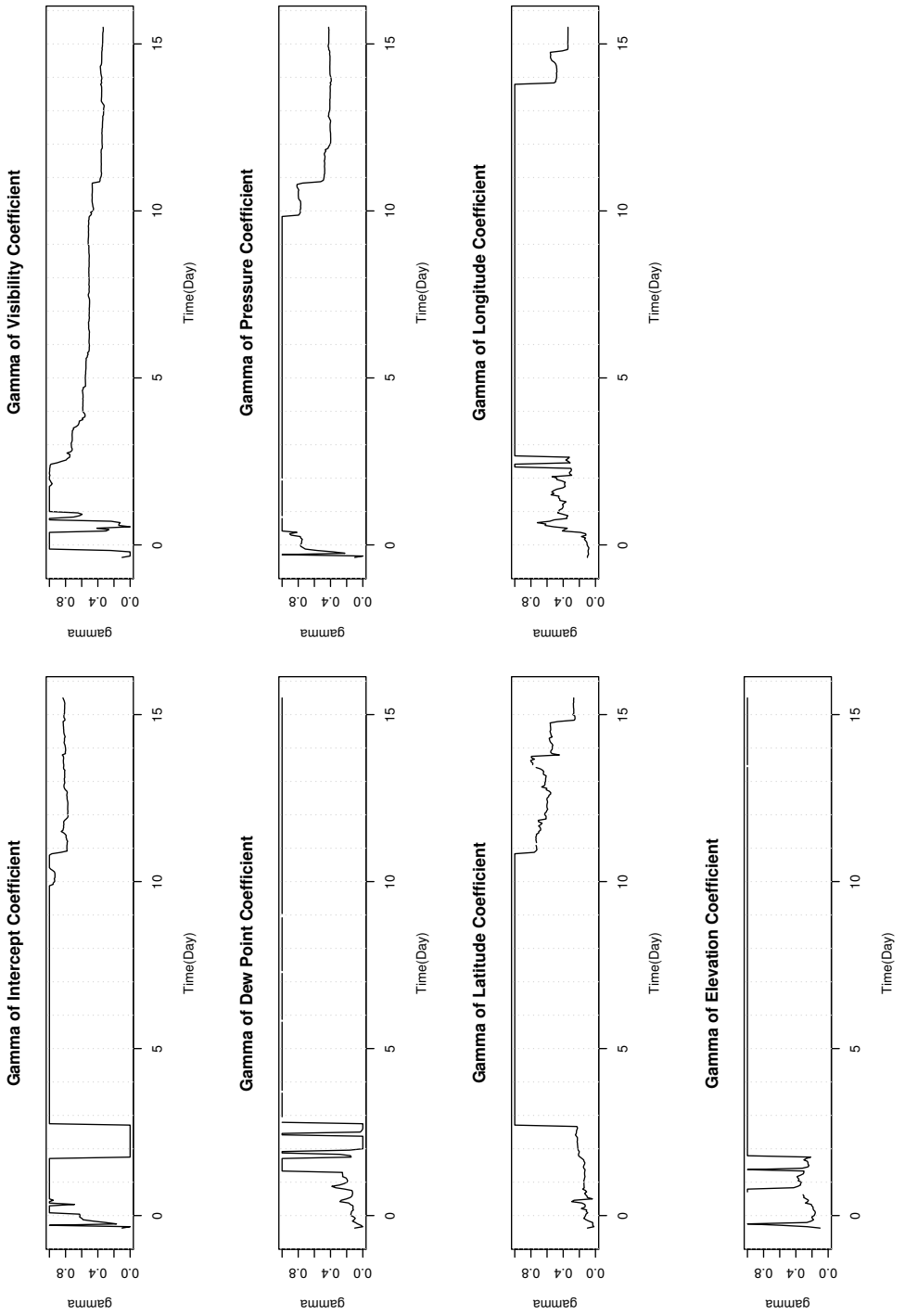

Figure 5.3: Different Coefficients Gamma with First Order Polynomial

CHAPTER 6

MODEL PERFORMANCE ANALYSIS AND EVALUATION

6.1 Result of Spatial Model Outlier Detection

We apply our model on weather variables: visibility, air temperature, dew point, and pressure from a dataset collected from sensors deployed at various locations in northeastern United State. Building model through our method is used for detecting the outliers. Those outliers could show up according to two main reasons. One is due to a drastic change in weather conditions, and the other is due to the erroneous reports by some sensors. According to our limitation of polynomials and residuals, we get our model and the normal state observations are lying within certain bounds of our model, which bounds could be set by ourselves depending on the customers requirement. In this way, we could easily find the outliers in the data.

Figure 6.1 shows the outlier detection performance of our scheme. The colored dots in the figure represent the sensor readings locations in the model applied to observed pressure values. The range of pressure is from low to high marked as red, orange and yellow. The dataset is still used those 90 weather sensor readings deployed in Northeast United State. As we known, there is a tornado passed from Westfield area (42.10N, 72.75W) to southwest Charlton (42.10N, 71.99W) in western Massachusetts on June 1, 2011. And we could observe that detection record this phenomenon with the sensor reading locating at (42.21N, 72.53W). This is the significant record, which could show the effectiveness of our model to detect outliers. Figure 6.2 is the effect spatial model of the pressure.

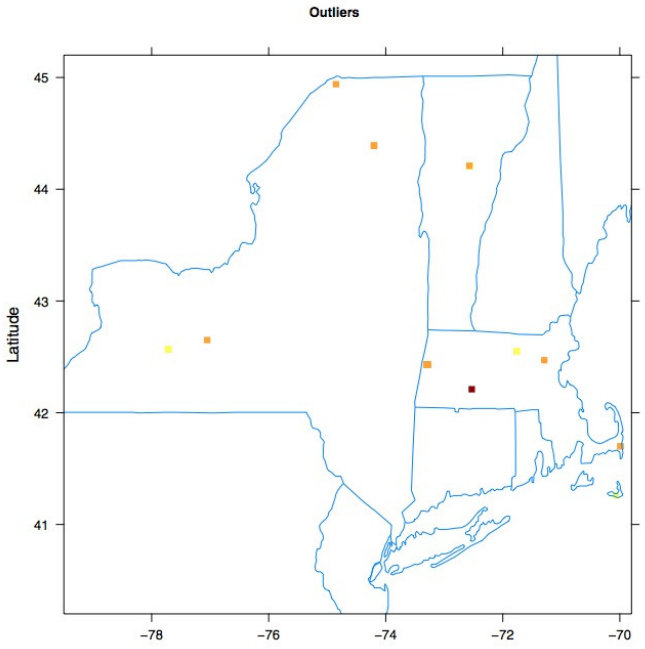

Figure 6.1: Pressure Outliers Detection for June 1, 2011 Tornado

6.2 Result of Temporal Extrapolation Model

We continue to apply our model on temperature temporal extrapolation to predict the next time slot value of sensor readings. Since each correlation between those variables should be stable in some degree, we are trying to figure out their regularity from the correlations between each of them that we calculated by our outlier detection system. If we could discover the regularity of correlation between two variables, then we could easily determine the expected values over time at each spatial coordinate. We use extrapolate temporally to predict sensor values for next several time steps. Here, we derive a mathematical method called Holt-Winters, which is also used for anomaly detection information to estimate network traffic volumes, since it could show the regularity of these correlationship among different sensor readings.

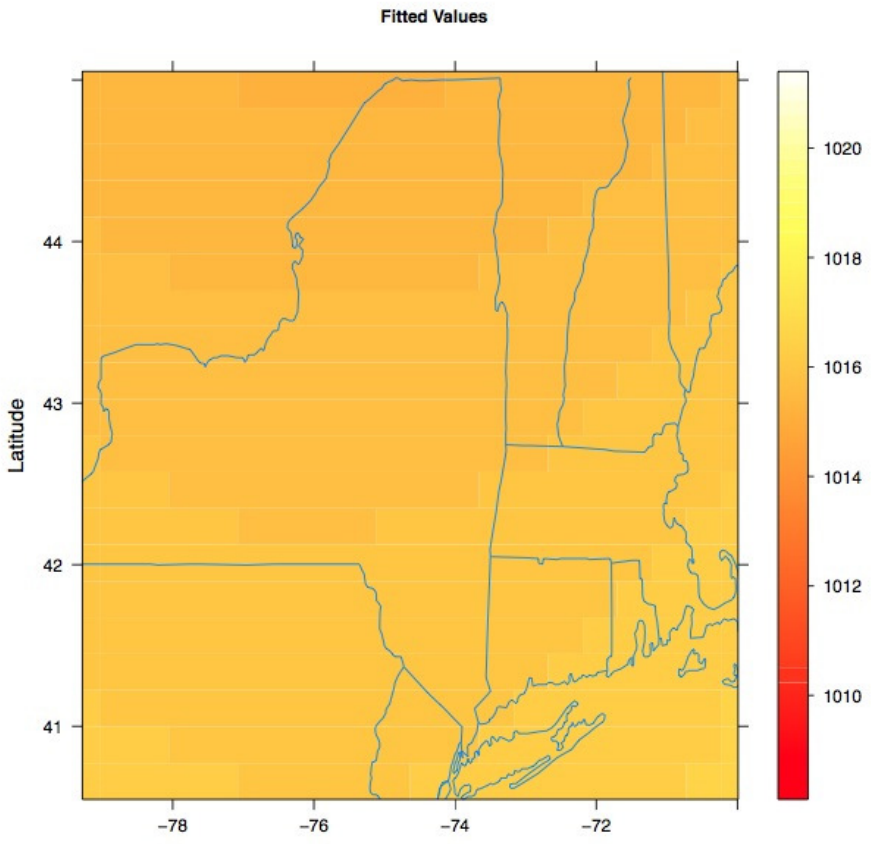

Figure 6.2: Temporal Extrapolation Graph For Pressure

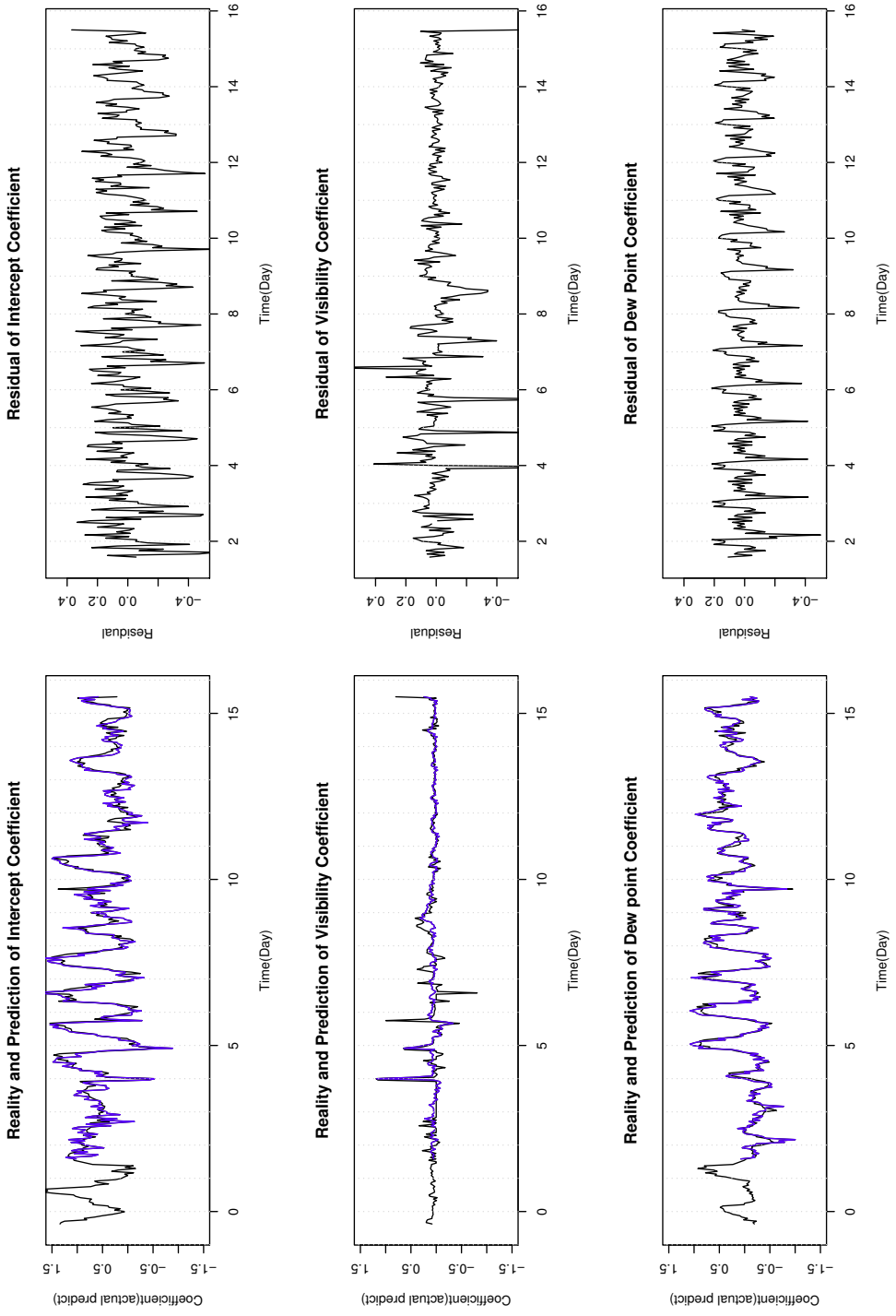

Figure 6.3: Coefficients Prediction with First Order Polynomial(1)

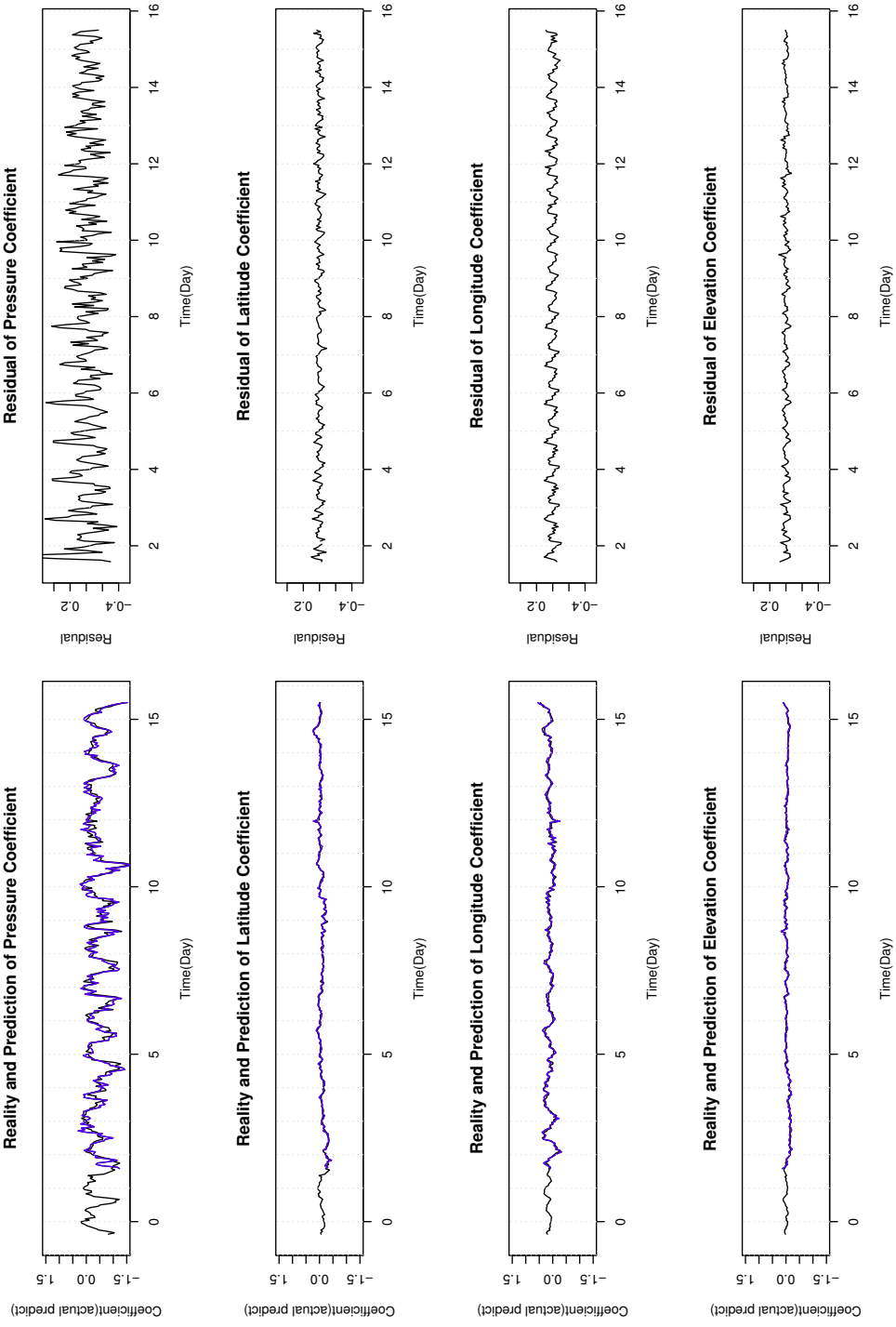

Figure 6.4: Coefficients Prediction with First Order Polynomial(2)

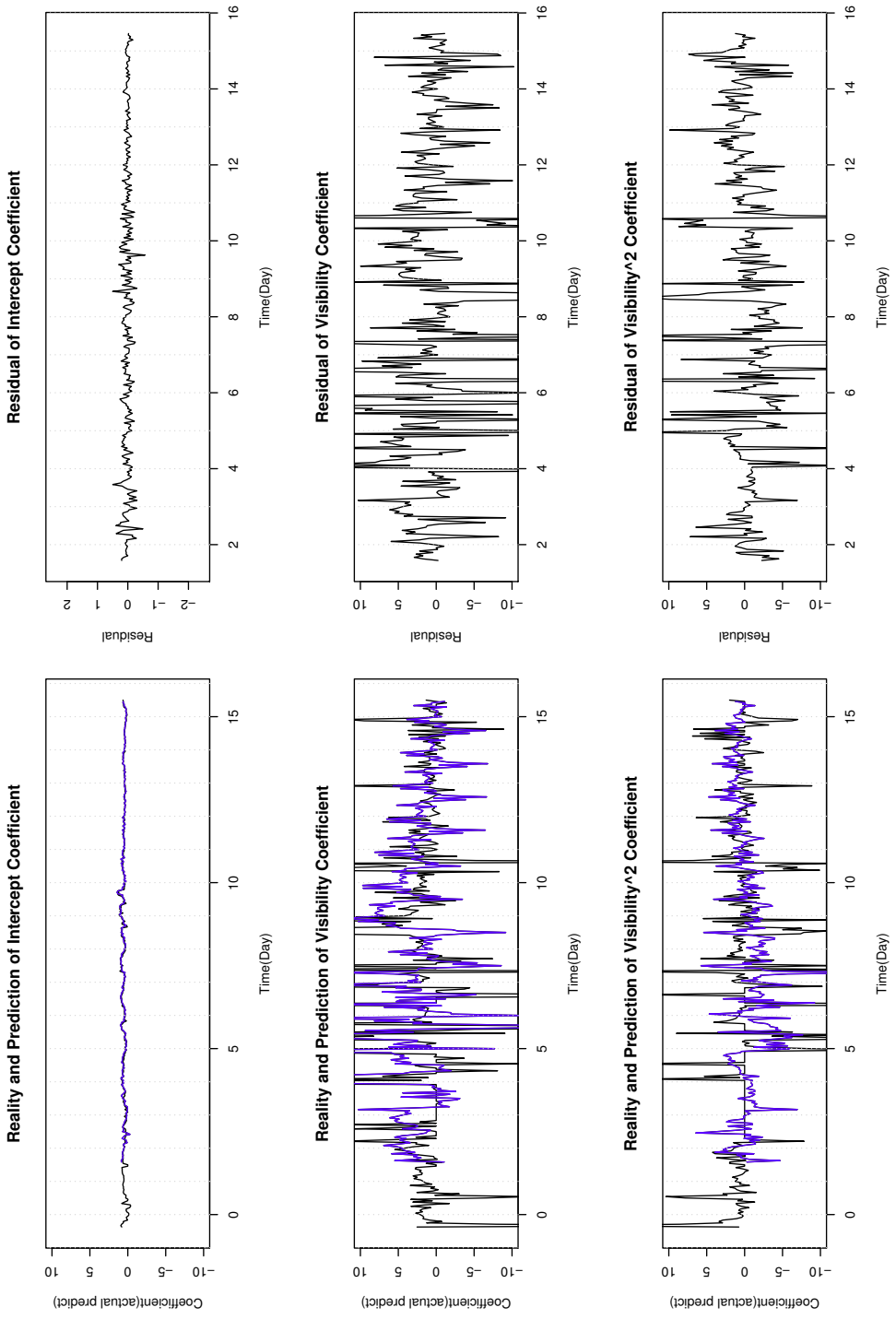

Figure 6.5: Coefficients Prediction with Second Order Polynomial(1)

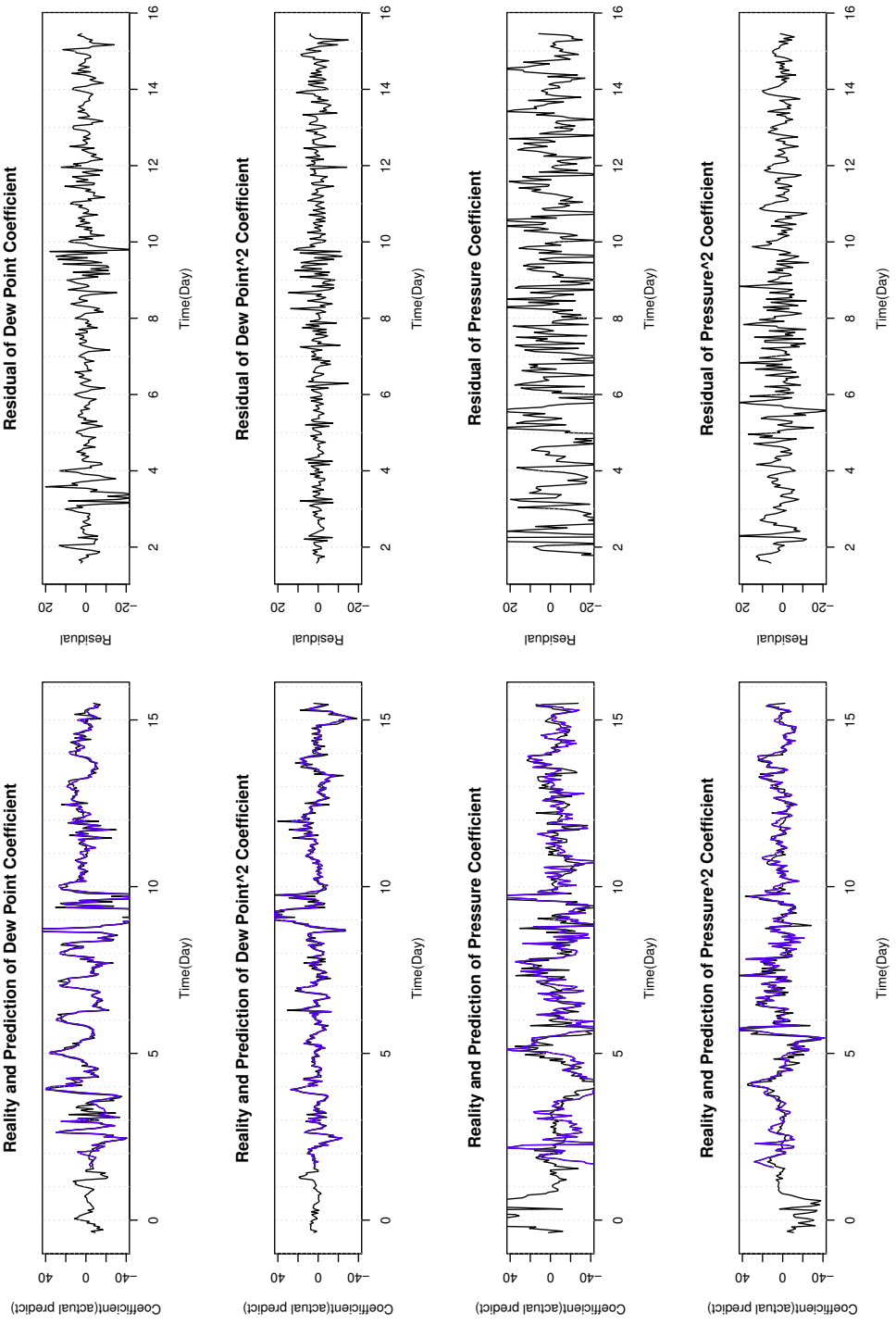

Figure 6.6: Coefficients Prediction with Second Order Polynomial(2)

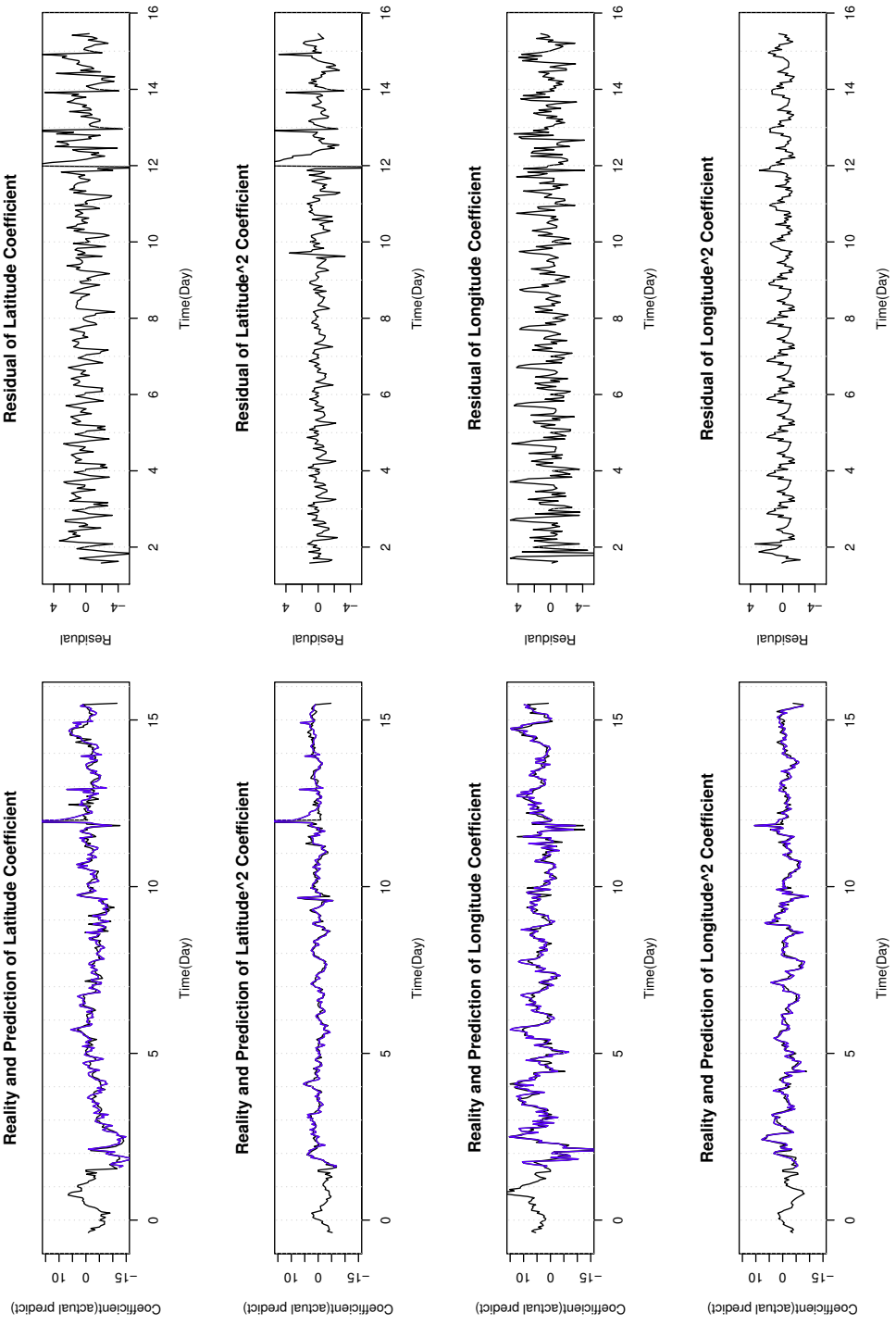

Figure 6.7: Coefficients Prediction with Second Order Polynomial(3)

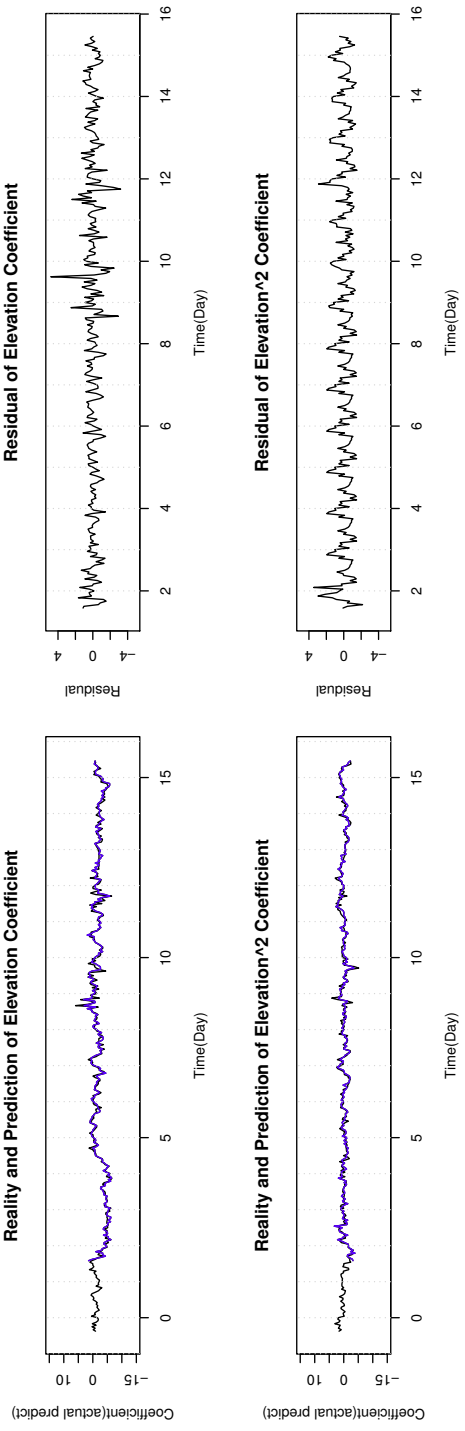

Figure 6.8: Coefficients Prediction with Second Order Polynomial(4)

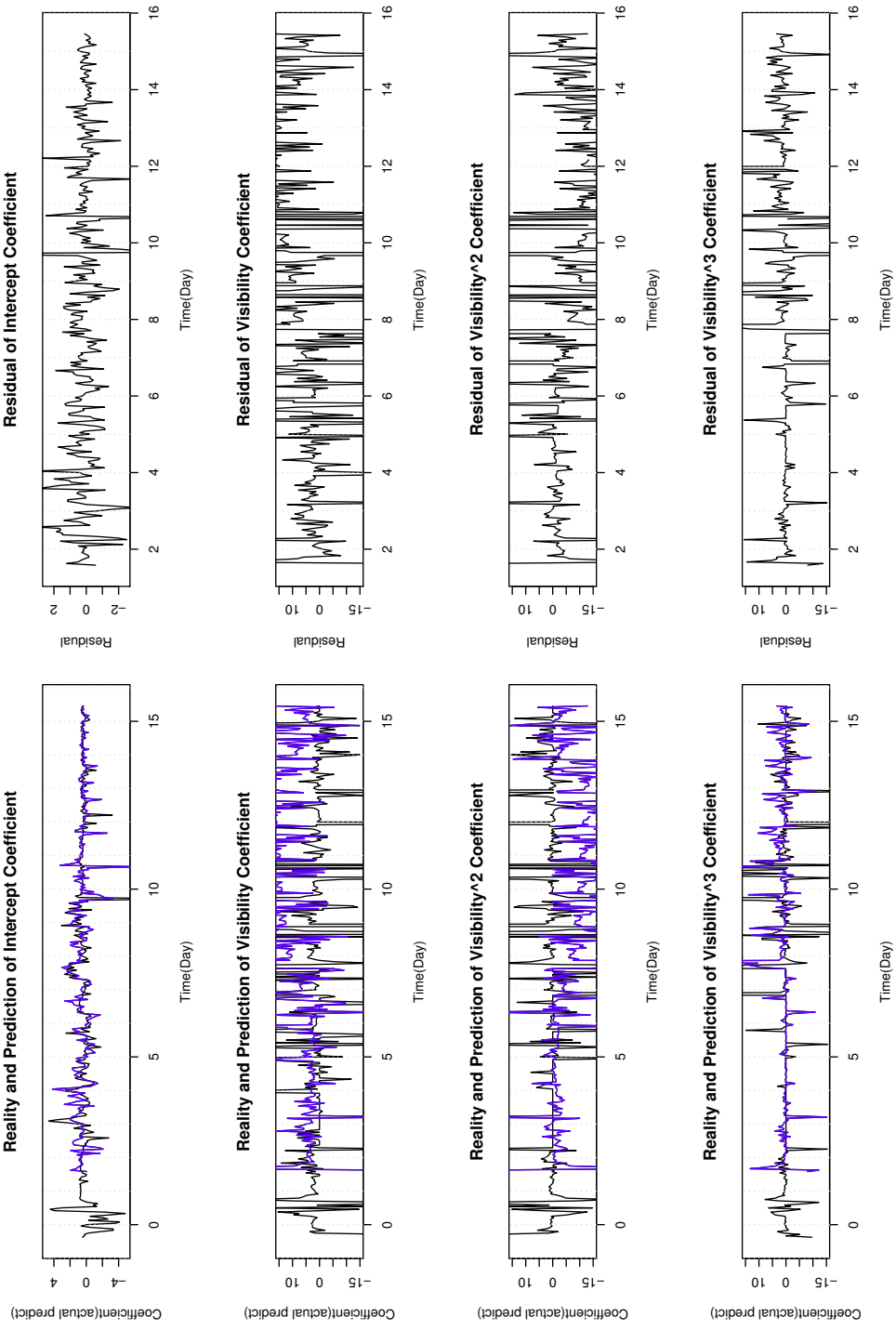

Figure 6.9: Coefficients Prediction with Third Order Polynomial(1)

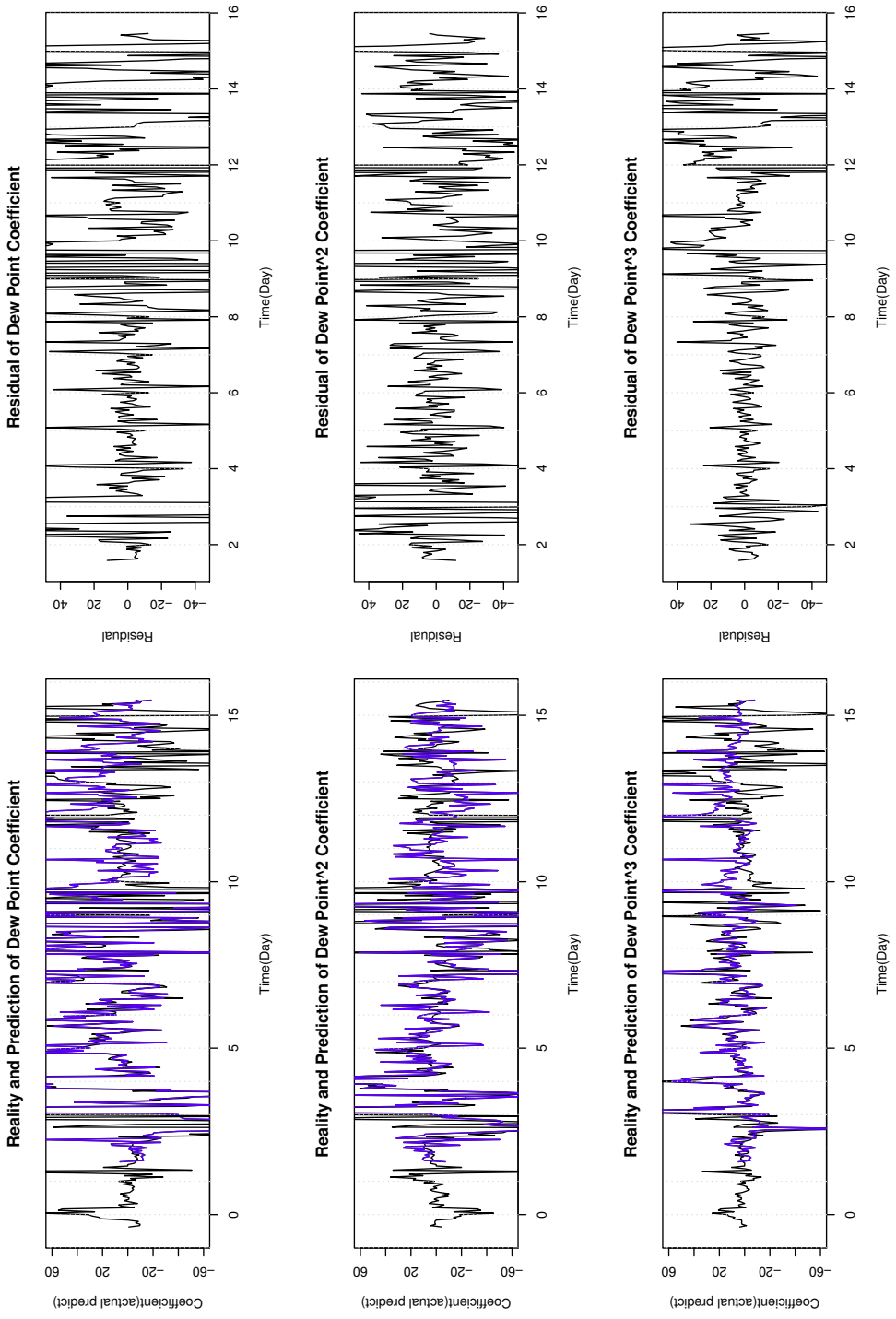

Figure 6.10: Coefficients Prediction with Third Order Polynomial(2)

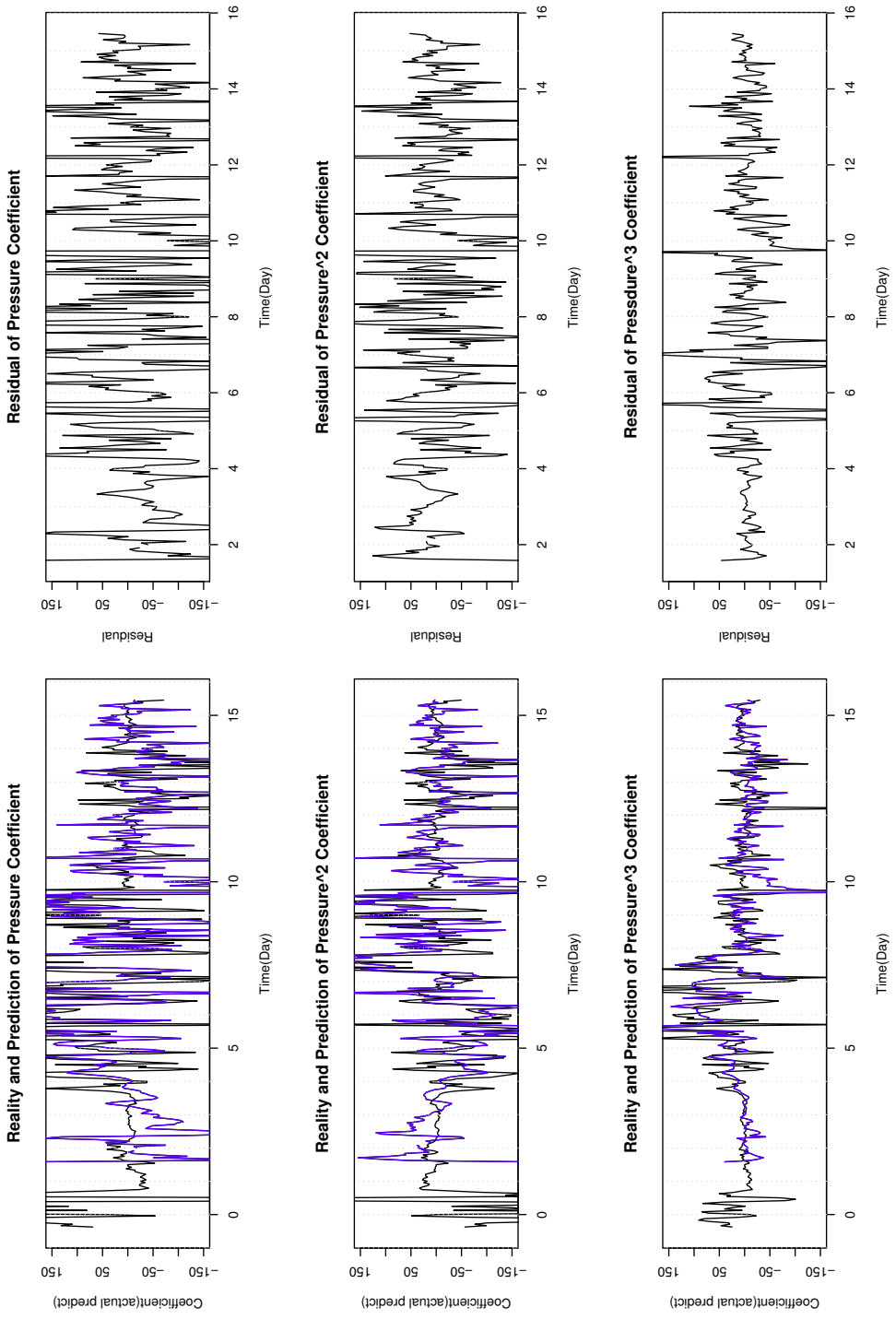

Figure 6.11: Coefficients Prediction with Third Order Polynomial(3)

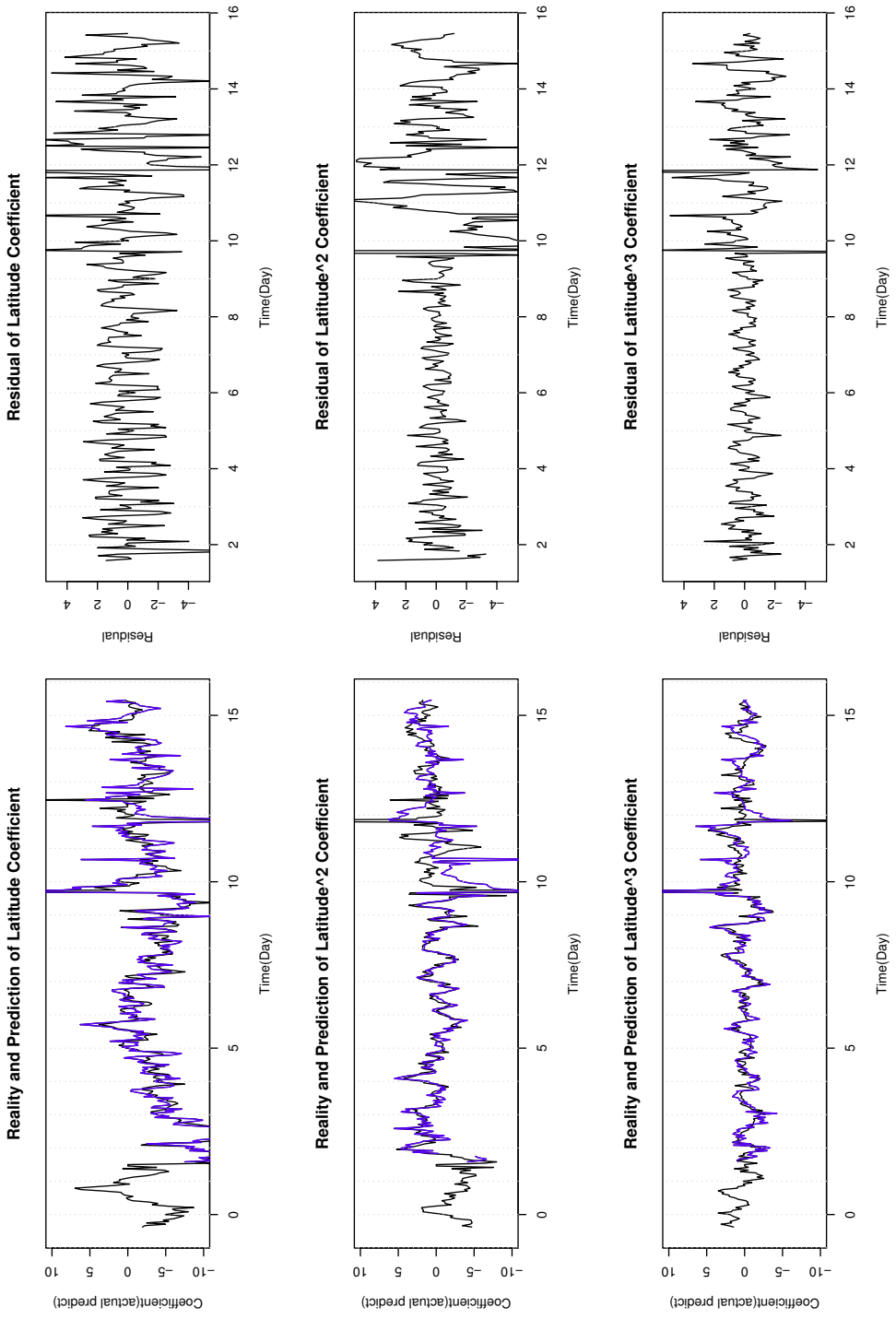

Figure 6.12: Coefficients Prediction with Third Order Polynomial(4)

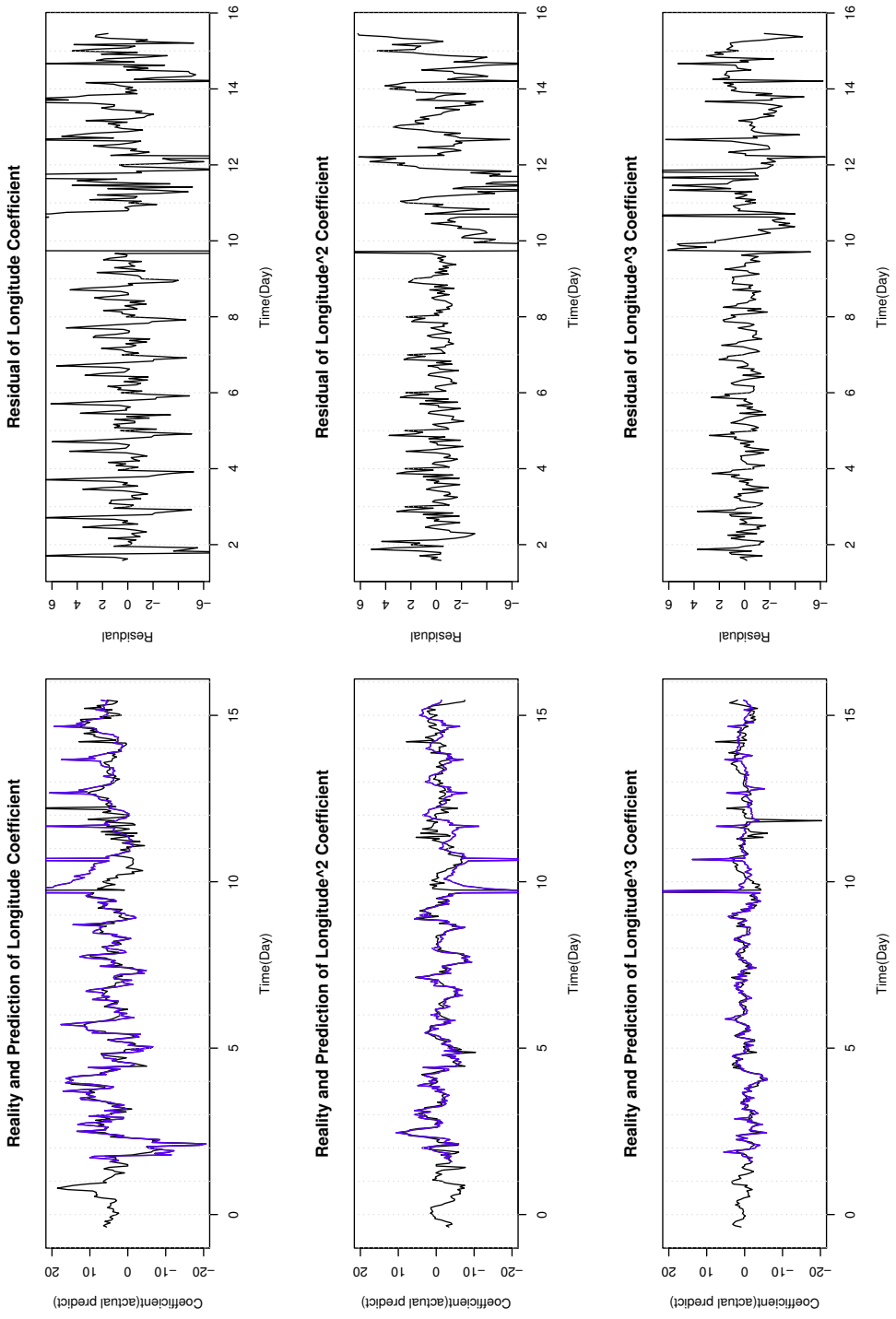

Figure 6.13: Coefficients Prediction with Third Order Polynomial(5)

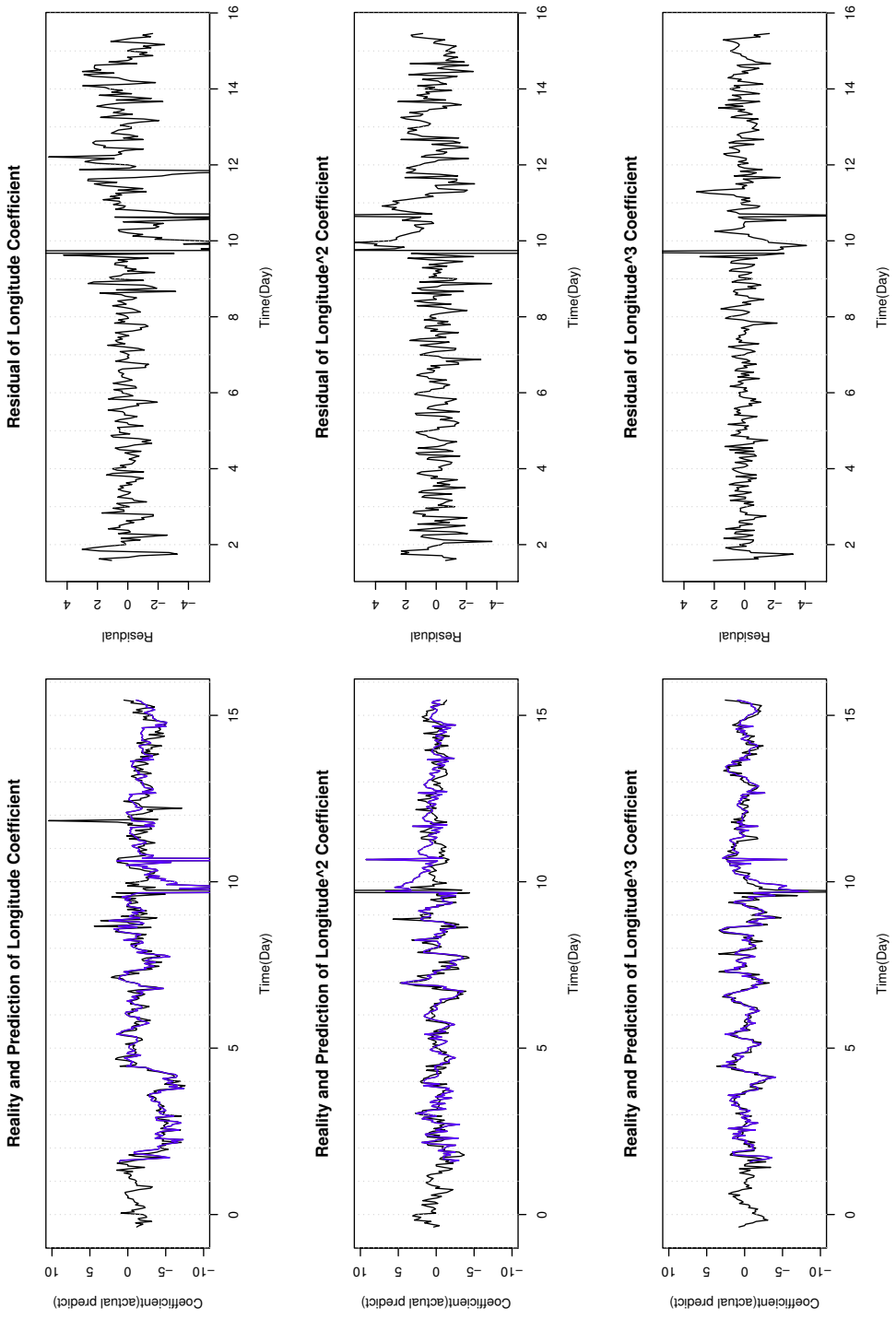

Figure 6.14: Coefficients Prediction with Third Order Polynomial(6)

We consider each one coefficient as a time series data. Figure 6.3 and figure 6.4 show first order of polynomial coefficients, and the performance of our prediction method. The black line is the reality coefficients according our selected model, blue line is predicted coefficient using an additive model with increasing or decreasing trend and seasonality. We can describe several useful conclusions. The prediction value is close to the reality value. We have the similar residual for each day, and the trend of residuals becomes downward. It proves that our method will get improved during a long period. And figure 6.5, figure 6.6, figure 6.7 and figure 6.8 show the coefficients with second order polynomials and the performance of our prediction. From these plots, we still could see that our prediction method with additive Holt-Winters method could effectively predict most of the coefficients. But compared with the first order polynomial, our residual gets larger. It might because the higher degrees of polynomial we have, the more coefficients we need to predict, it means the more slight errors we have to consider. Then, we continue applying our method to the model with three degree of polynomial and check the performance of our method. Those following six plots (figure 6.9, 6.10, 6.11, 6.12, 6.13,and 6.14) are the coefficients of three degree of polynomials in our model. These plots show that, the coefficients with high degree of polynomials do provide larger standard error. So, our performance becomes worse than prediction with first order polynomial and second order polynomials.

Figure 6.15 shows the example of fitting different degrees of polynomials to estimate the temperature values with the residuals. While increasing degrees of polynomials, it is evident that the residuals for the model become worse. It is because we need to predict more coefficients with high polynomials and sum up all the variables. So we use first order to do the temporal extrapolation with small residual. And our model become better as the experiment time goes long since the residuals become smaller and smaller.

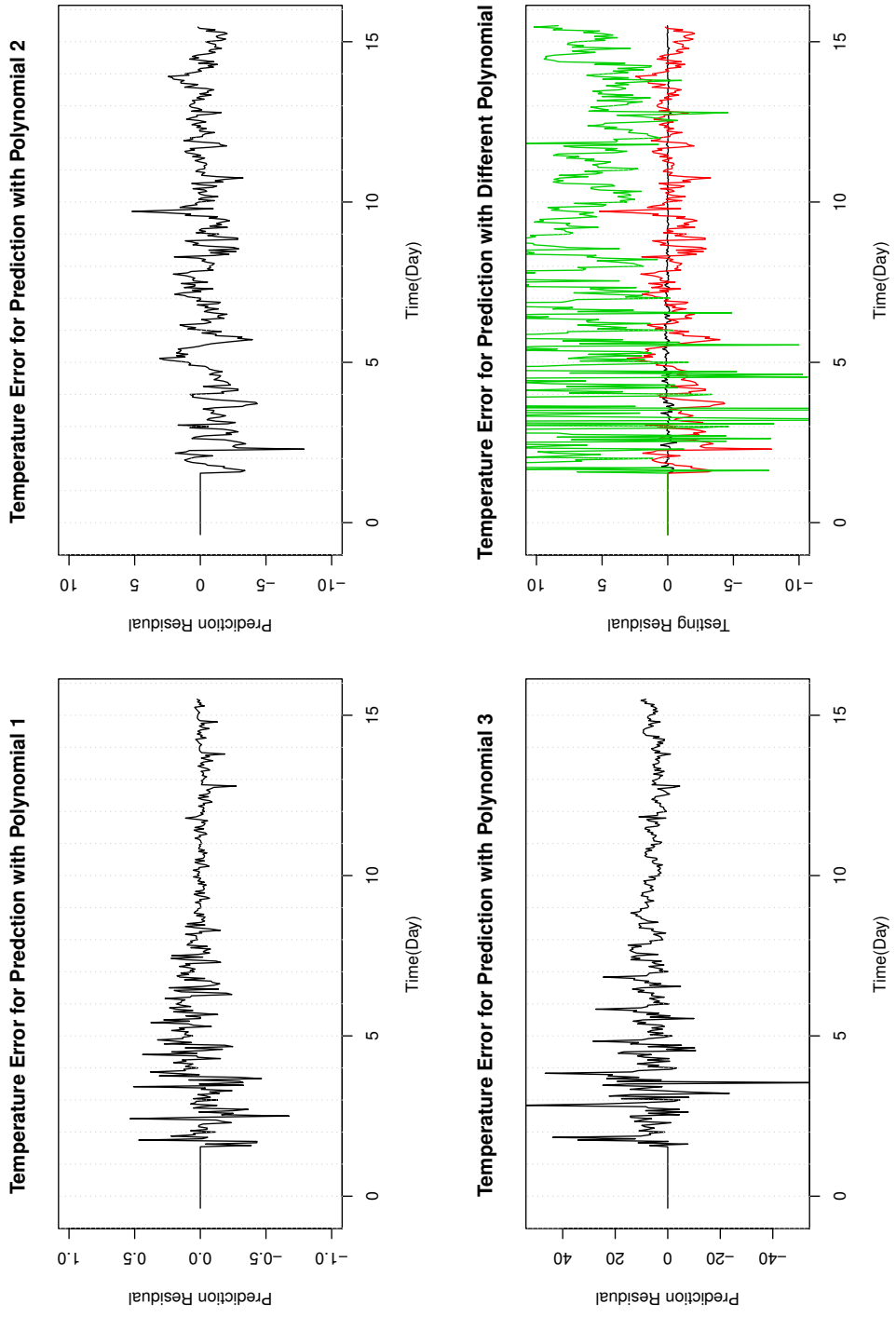

Figure 6.15: Residual for Prediction with Different Polynomials

6.3 Evaluation of Temporal Model

To prove our temporal model selection effective as described above, which could predict sensor values for next several time steps, we are trying to verify our predictions against real weather prediction model. One thing should be mentioned is that this prediction is not equivalent to weather forecasting, but predict sensor values crossing different domains.

At this time, we collect five weather variables, additional variable is humidity to prove our temporal model mutable and widely used. After two periods of model formed, our temporal model selection scheme could predict next time step model automatically. In the other word, It could provide the next time step sensor values according to other domain sensors' values.

Figure 6.16 shows the following one step prediction proformance in short term, while our scheme just formed. The x axis is the actual value of temperature. The y aixs is the predict value of temperature. The black circles in the figure represent the temperature from real weather prediction model at one time step in different sensor stations. The blue crosses represent the temperature from our temporal model. This figure clearly shows the effectiveness of our model to predict sensor value in the data, especially when it becomes as the same accurate as real weather model.

Figure 6.17 is also a following one step prediction proformance at 16 on Dec. 16, 2012, after our temporal model formed one month later. Our prediction values still have a great performance.

Table 6.1: Percentage of Mean Squared Error

Mean Squared Error	*Percentage*
0<Mean Squared Error $\leq$ 1	65.8%
1< Mean Squared Error $\leq$ 4	23.6%
4< Mean Squared Error $\leq$ 9	9.5%
Residual > 9	1.1%

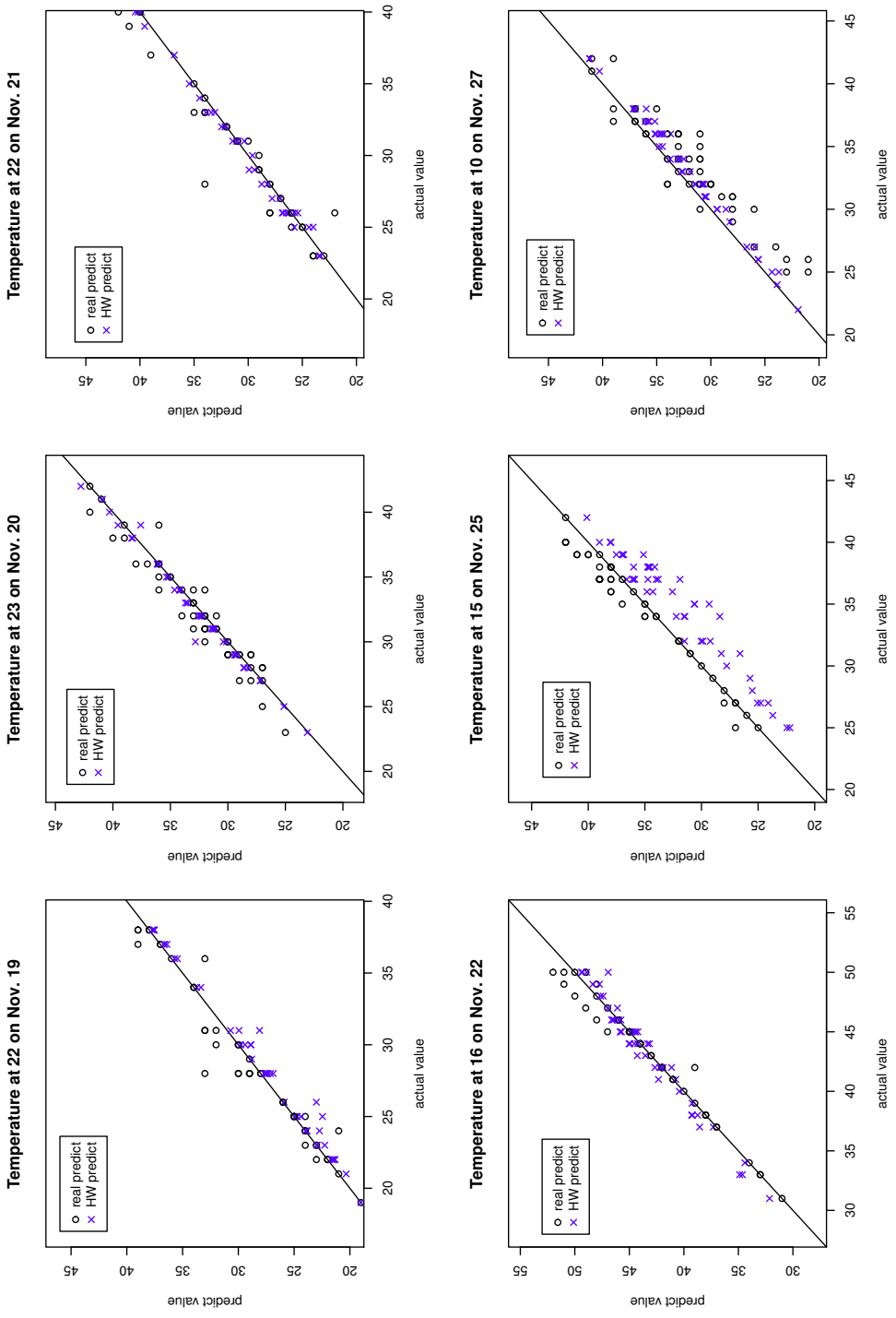

Figure 6.16: Temporal Extrapolation For Temperature In Short Term

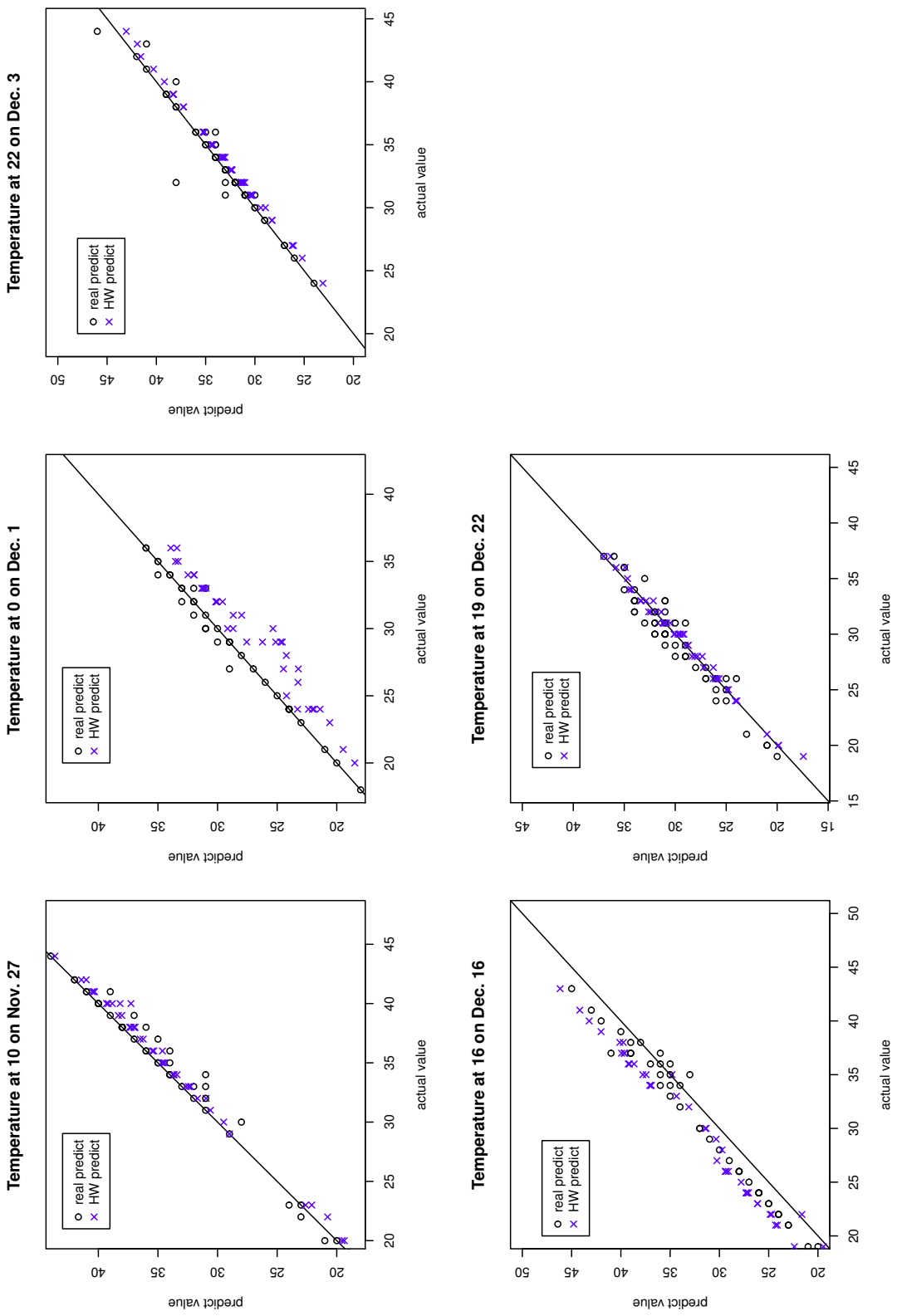

Figure 6.17: Temporal Extrapolation For Temperature one month later

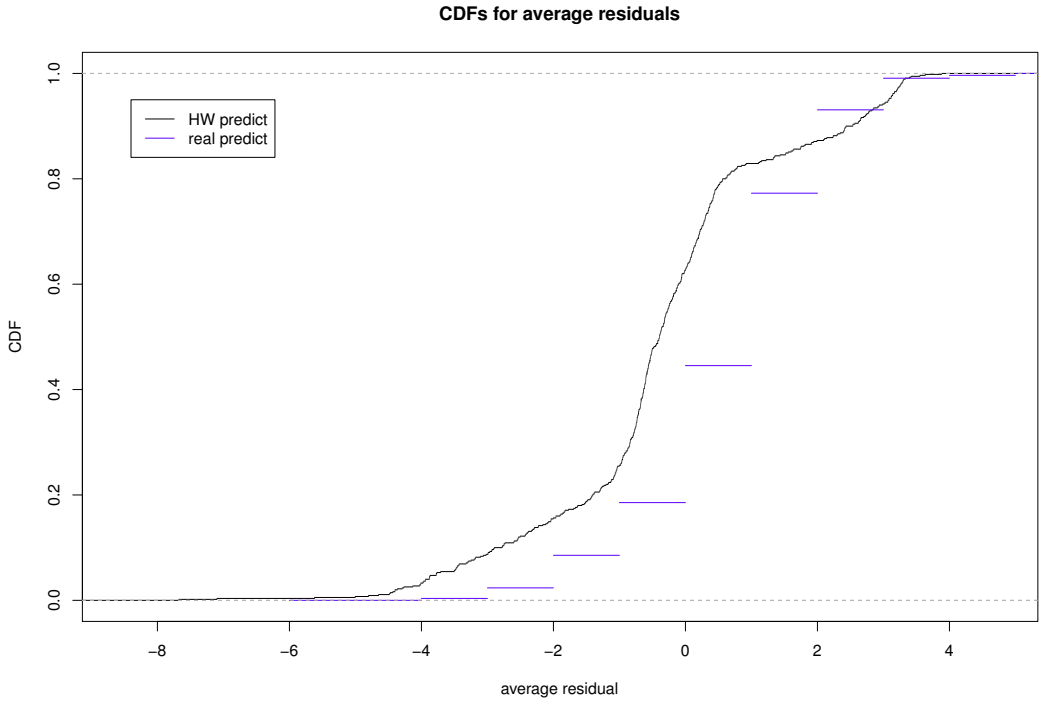

Figure 6.18: CDFs of Average Residual

To further demonstrate this conclusion, we observe the mean squared error between temporal model values and actual values for each time step in Table 6.1. Total number of temporal model is 918, 65.8% temporal model of mean squared error is less than 1 degree, 23.6% temporal model of mean squared error is between 1 and 2. So, there is 89.4% of models could provide high accurate and good performance. Meanwhile, the mean squared error between real weather prediction model and actual values is 1.36 degree and the mean squared error of our model is -0.544. At the same time, the CDFs of Holt-Winters prediction and real weather prediction in Figure 6.18, and the time series plot of mean squared error (round to integer) with time series in Figure 6.19, all of these also prove that our Holt-Winter extrapolation model could provide as the same great performance as the accurate sensor values.

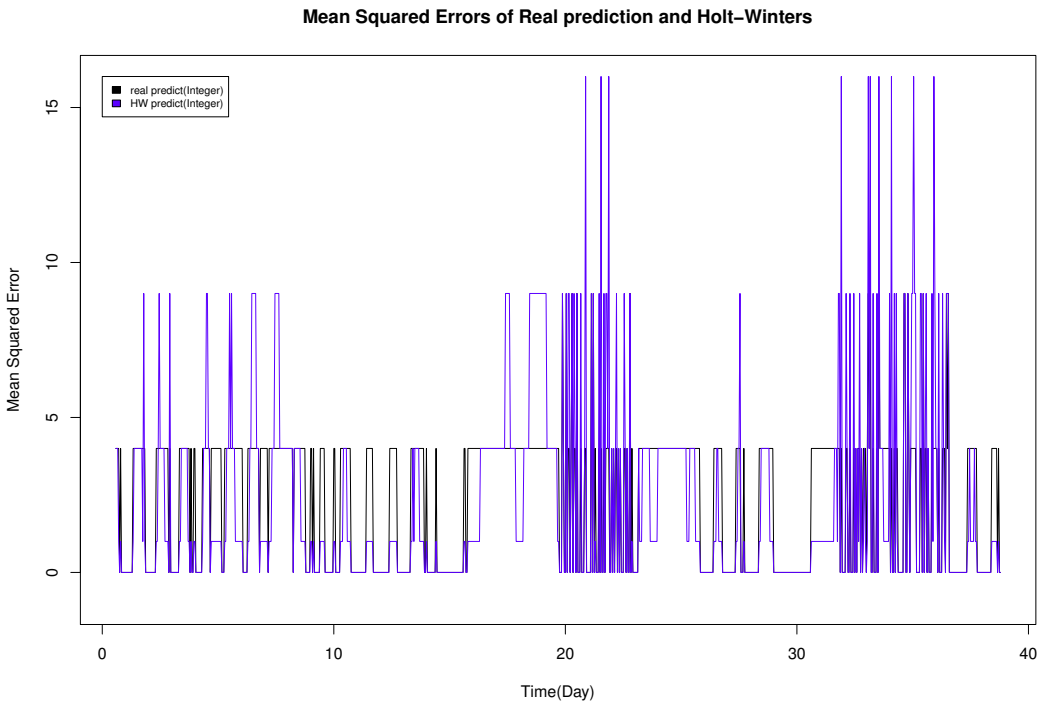

Figure 6.19: Mean Squared Error of Real Prediction and HW Prediciton

CHAPTER 7

CONCLUSION

In this paper, we have analyzed several sensors data relationship, implemented a method for spatial model and temporal model to detect sensor values no matter at different spatial coordinates at an instant time or detect the expected value of sensors at all coordinates in next several time step. Our spatial model is modeling of different domain sensor data based on multiple regression, and temporal model utilizes exponential smoothing to evaluate next time step model. Our temporal model shows promise in detecting next time step sensor data values, provides great performance in evaluating next one time step model in weather sensing application, and as the same as real weather prediction model, especially under the situation that add additional sensor data, then re-do the prediction. It could illustrate that our model is not domain specific, and could be applied in any application domains with continuous sensor data of Cyber-Physical System. We do believe our research could help setting up a scalable deployment of Cyber-Physical System.

BIBLIOGRAPHY

[1] C.-W. Ten, C.-C. Liu, and Manimaran, G. Vulnerability assessment of cybersecurity for scada systems. *Power Systems, IEEE Transactions*, 4 (2008).

[2] Cheng, A. M. K. Cyber-physical medical and medication systems.

[3] Cyril Voyant, Marc Muselli, Christophe Paoli Marie-Laure Nivet. Numerical weather prediction (nwp) and hybrid arma/ann model to predict global radiation. *Energy* (2012).

[4] H. F. Wedde, S. Lehnhoff, C. Rehtanz, and O.Krause. Distributed embedded real-time systems and beyond: A vision of future road vehicle management. *34th Euromicro Conference Software Engineering and Advanced Applications*, 3–5 (2008).

[5] J.F.Ojo. On the performance and estimation of spectral and bispectral analysis of time series data. *Asian Journal of Mathematics and Statistics* (2008).

[6] L. Sha, S. Gopalakrishnan, X. Liu, and Wang, Q. Cyber-physical systems: A new frontier.

[7] Lee, Edward A. Cps foundations.

[8] M. Ilic, L. Xie, U. Khan, and Moura, J. Modeling future cyber- physical energy systems. *Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, IEEE*, 19 (2008).

[9] Ma, Q., and Steenkiste, P. A weather forecasting system using concept of soft computing: A new approach. In *International Conference on Advanced Computing and Communications* (2006), pp. 353–356.

[10] Paras, Sanjay Mathur, Avinash Kumar, and Chandra, Mahesh. A feature based neural network model for weather forecasting. *World Academy of Science, Engineering and Technology* (2007).

[11] Work, Daniel B., and Bayen, Alexandre M. Impacts of the mobile internet on transportation cyberphysical systems: Trafc monitoring using smartphones.

[12] Y. Gu, A. McCallum, and Towsley, D. Detecting anomalies in network traffic using maximum entropy estimation, Oct. 2005.