Spring 2015

Uncertainty quantification and calibration of physical models

Xian He
Purdue University

Graduate School Form
(Revised /)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Xian He

Entitled

UNCERTAINTY QUANTIFICATION AND CALIBRATION OF PHYSICAL MODELS

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Hao Zhang

Qianlai Zhuang

Bruce Craig

Bo Li

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Hao Zhang

Approved by Major Professor(s): _____

_____

Approved by: Jun Xie 03/03/2015

Head of the Graduate Program Date

UNCERTAINTY QUANTIFICATION AND CALIBRATION OF PHYSICAL MODELS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Xian He

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2015

Purdue University

West Lafayette, Indiana

To my dear family.

ACKNOWLEDGMENTS

I am thankful to NSF Division of Information and Intelligent Systems (IIS-1028291), National Science Foundation (NSF- 1028291 and NSF- 0919331), NSF Carbon and Water in the Earth Program (NSF-0630319) for the financial support I had, that gave me the opportunity to focus on my research and travel to conferences.

I am most thankful to my advisor, Prof. Hao Zhang, for all the advice in the last five years, which has changed my perspective and my attitude to research and life. His patience, wisdom and kindness have made my graduate study a very valuable and rewarding experience. He has also provided many opportunities for me to collaborate with scientists in other departments, where I enhanced my skills in communication and collaboration. Prof. Zhang is always available to help his students with any questions regarding academic, career and life. I really appreciate his support and advising which always drives me to be more excellent.

I am also very fortunate to work with Prof. Qianlai Zhuang. The opportunity to collaborate with his team is a very unique one for me to apply statistics to environmental science. He provided outstanding resources for me to learn the science, and his door is always open for any questions. The collaboration with his lab researchers–Dr. Min Chen, Dr. Xudong Zhu, Shaoqing Liu and other graduate students is a very productive and inspiring research experience. I am very thankful for his mentoring and support.

Many thanks go to Professor Bruce Craig and Prof. Bo Li for their willingness to serve on my committee and helpful suggestions to my work. During my working experience in statistical consulting service, I have learned a lot from Prof. Bruce Craig and developed my statistical consulting skills. Prof. Bo Li is always very gentle and willing to help me in all aspects of graduate life.

I would like to acknowledge current and past members of Prof. Hao Zhang's research group, including Cheng Liu, Yong Wang, Juan Hu, Inkyung Choi, Yen-ning Huang, Whitney Huang, Piyas Chakraborty and Kelly-Ann Dixon Hamil, for discussing research questions over the past five years and providing valuable comments.

The Department of Statistics at Purdue has provided a wonderful environment for graduate students to learn, teach and carry out research. I had the opportunity to be an instructor to teach STAT301 and STAT225. From this valuable experience, I learned how to be a good instructor, convey concepts and ideas. More importantly, as an international student, I gained the confidence in public speaking in English. I would like to thank the staff members, Doug Crabill for his help with any computational issues, Teena Erwin and Becca Miller for their assistance with any questions during my graduate life, for their dedication to providing outstanding resources to the department.

Finally, to my dear family, there are no words that can truly express the level of gratitude and appreciation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

TEM	Terrestrial Ecosystem Model
C	Carbon
N	Nitrogen
GPP	gross primary production
NPP	net primary production
NEP	net ecosystem production
GP	Gaussian process
MVN	Multivariate Normal Distribution
PAR(p)	Periodic autogression with order p
PIAR(p)	Periodic integration autogression with order p
VAR(p)	Vector autogression with order p
$\mathcal{N}_{n,q}$	matrix normal distribution with dimension (n, q)
$\mathbb{R}_{n,q}$	matrix with dimension (n, q)
LHS	Latin Hypercube Sampling

ABSTRACT

He, Xian Ph.D., Purdue University, May 2015. Uncertainty Quantification and Calibration of Physical Models. Major Professor: Hao Zhang.

An *ecosystem model* is a representation of a real complex ecological system, and is usually described by sophisticated mathematical models. Terrestrial Ecosystem Model (TEM) is one of the ecosystem models, that describes the dynamics of carbon, nitrogen, water and other vegetation related variables. There are uncertainties in the TEM which are attributed to inaccurate input data, insufficient knowledge of the parameters, inherent randomness and simplification of the physical model. Quantification of uncertainty of such an ecosystem model is computationally very heavy. Bayesian calibration method has been used as an efficient way to calibrate and quantify uncertainties of the computer models.

In this work, I develop a new approach to emulate the TEM, and to estimate the parameters along with associated uncertainties. TEM has been implemented as a deterministic computer code model. In this computer model, the inputs are environmental variables and underlying parameters, and the outputs are gross primary production (GPP), net ecosystem production (NEP) and other variables. To make predictions of future outputs from the computer model, I also estimate the underlying parameters. With an efficient Bayesian approximation, statistical models are developed to obtain inference for the parameters and then make predictions at future time point.

Chapter 1 is an introduction to the research problems. In Chapter 2, I discuss the uncertainty arose from temporal scales. In Chapter 3, I discuss the Bayesian uncertainty quantification method and further developed Bayesian calibration of parameters with application to TEM in Chapter 4.

1. INTRODUCTION

1.1 Motivation

An *ecosystem model* is a representation of the real complex ecological system, and is usually described by sophisticated mathematical models. Terrestrial Ecosystem Model (TEM) is one of the complex ecosystem models. It describes the dynamics of carbon, nitrogen, water and other vegetation related variables using environment variables such as temperature, precipitation, global radiation and carbon dioxide to make estimation of carbon and nitrogen fluxes. Gross primary production (GPP) and net primary production (NPP) are the two important carbon fluxes modeled by TEM. GPP is the total amount of energy primarily produced by plants through photosynthesis. Part of this energy is used by plants for respiration and maintenance of existing tissues. The remaining energy is referred to as NPP, which is the entire amount of energy produced minus the respiration by plants. Some NPP goes toward growth and reproduction of primary producers, while some is consumed by herbivores. NPP is a measure of the plant growth, which is an important reflection of the global climate change. So the estimation of NPP is very essential to the entire carbon dioxide exchange in ecosystem. It would reflect the assessment of pollution levels and influence potential regulatory policies.

Many process-based computer code models have been developed based on the differential equations of TEM [Raich and Schlesinger, 1992]. However, due to inaccurate inputs, insufficient knowledge of the parameters in the model and inherent randomness of the system, the model output has inherited uncertainty. Quantifying the uncertainty in the physical model is very important in forecasting and assessing the variability of outputs from the system. Most previous work focused on the estimation of carbon and nitrogen fluxes using process-based environmental models, while

very few work discussed about the uncertainty quantification. Uncertainty analysis is the study about how the distribution of the outputs depends on the inputs and parameters. The quantification of uncertainty provides us the confidence level in the estimation of outputs and how robust the conclusion of the model results. Also, we could assess the efficiency of various models based on their corresponding uncertainty levels and decide weight on different models. Further study of how much uncertainty might be induced by learning about some specific inputs is called a *sensitivity analysis*. It tells us the source of the uncertainties and what are more important to know. [Zhuang et al., 2009] estimated NPP and their associated uncertainties using geospatial statistical approaches.

It is widely recognized that accurate quantification of carbon fluxes (e.g., GPP, NPP and NEP) is becoming increasingly important both scientifically and economically. With the development of satellite and remote sensing technology, it is feasible to get massive and finer temporal and spatial resolution data. Then, how to improve model parametrization with these massive data? How to quantify uncertainties from various sources? This dissertation will address these questions. In next section, I will discuss the sources of uncertainties in ecosystem models.

1.2 Background

1.2.1 Sources of Uncertainties

Uncertainties exist in observations and physically-based ecosystem models, and can be caused by many factors, such as measurement error, model simplification, inherent randomness of the system, etc. Uncertainty analysis is to assess the distribution of output induced by distribution of inputs. In this work, I focus on three important sources of uncertainties: the temporal scale variation, model inadequacy and parameter uncertainty in the model.

(1) *Scale Variation*

With different choices of spatial or temporal scales, the input variables will be different, and there will be uncertainties in the model output. Many statistical models for environmental studies can be run at different scales, e.g., daily, weekly or monthly data. It is important to know when and how these models of different scales differ. Although there are some empirical studies on models of different scales ([Berrocal et al., 2012, Mueller et al., 2010, Patil and Deng, 2012]), there is a lack of theoretical discussion and explicit conclusions on the scaling problem.

In many environmental studies, choosing a suitable temporal scale (e.g, hourly, daily, weekly or monthly) is one of the most important steps. With the improvement of remote sensing technology, it is feasible to acquire data at various spatial and temporal resolutions. We can therefore run a model at a larger scale or run it at a finer scale and then upscale the results. How would the results differ? I will discuss this problem in Chapter 2.

(2) *Model Inadequacy*

The real ecosystem is so complex that any physical model or mathematical model only reflects the current knowledge guided by observations. While the data have limitations and our knowledge is limited, how to quantify the uncertainties of model inadequacy? In Chapter 3, I will quantify this type of uncertainty using Bayesian framework with monthly temporal scale. Then, the uncertainty is obtained by comparing the real observations with the statistical models.

(3) *Parametric Uncertainty*

Physical models, which are implemented as computer code models, have many input parameters which are chosen in advance for the system. The computer models are deterministic, while the real systems are random. In the computer model, there are two types of inputs: variable inputs (temperature, precipitation and etc.) and unknown parameters. Outputs are unknown functions of inputs.

Some of the parameters are very sensitive that which value to choose makes big difference in the behavior of the process. Especially when the parameters are continuous. Uncertainty quantification and calibration of those parameters are of vital importance in developing computer code models. Hence it is always one of the essential problems for environmental scientists to calibrate the models.

1.2.2 Calibration

Calibration is the process of adjusting the parameters until the model outputs fit the observations. It is the very first step for any further application of the physical models. Traditional method for calibration is to manually adjust the parameters by comparing some metrics from the physical model and the observations (e.g., [Chen and Zhuang, 2012];[Zhu and Zhuang, 2013]). This approach is very computationally expensive. [Kennedy and O'Hagan, 2001] thoroughly describes the Bayesian calibration approach on unknown calibration parameters. This work considered all sources of uncertainties and improved the discrepancy between the model predictions and observations. [Tang and Zhuang, 2009] applied the Bayesian inference approach on the calibration of TEM.

Under the Bayesian framework of [Kennedy and O'Hagan, 2001], the computer code model of TEM could be approximated by Gaussian process in which the computer model is viewed as random process. By comparing the computer model and real observations, the parameters are estimated to minimize the discrepancy between them. However, to estimate the parameters in dynamic ecosystem model where the output is time dependent, we need to improve the existing Bayesian calibration approach to handle the time series inputs and outputs. Inspired by the approaches in [Chen and Zhuang, 2012] as well as [Kennedy and O'Hagan, 2001], I developed a new Bayesian calibration method on dynamic ecosystem models. In this new approach, the first step is to develop an efficient emulator to represent the computer model. Kriging and CoKrging are two different ways to develop the emulator. In these two

ways, estimation of parameters in covariance functions is one essential step. Next, I will discuss different covariance functions.

1.2.3 Covariance Functions

Covariance function describes the spatial covariance of a random process. In spatial statistics, it is defined as

$$C(\boldsymbol{s},\boldsymbol{t}) = cov(\boldsymbol{y}(\boldsymbol{s}),\boldsymbol{y}(\boldsymbol{t})) \tag{1.1}$$

where $\boldsymbol{s}$ and $\boldsymbol{t}$ are different two locations. $\boldsymbol{y}(\boldsymbol{s})$ and $\boldsymbol{y}(\boldsymbol{t})$ are the corresponding random variables of these two locations. A second-order stationary(SOS) process is a process which satisfies the following two conditions

$$\begin{aligned} E(\boldsymbol{y}(\boldsymbol{s})) &= constant, \boldsymbol{s} \in \mathcal{R}^d \\ C(\boldsymbol{s},\boldsymbol{t}) &= C(0, \boldsymbol{t}-\boldsymbol{s}) \end{aligned}$$

The covariance function of SOS only depends on the distance $\boldsymbol{h} = \boldsymbol{t} - \boldsymbol{s}$. Furthermore, if a stationary covariance function only depends on the norm $||\boldsymbol{t} - \boldsymbol{s}||$, it is called *isotropic*. There are a few parametric families of covariance functions which are frequently used in literatures. Here, we will introduce them. For $\boldsymbol{h} \in R^d$, any $d > 0$,

(1) Powered Exponential Covariogram

$$C(\boldsymbol{h}) = \theta_1 exp(-(\|\boldsymbol{h}\|/\theta_2)^\alpha), \theta_1 > 0, \theta_2 > 0, 0 < \alpha \leq 2 \tag{1.2}$$

when $\theta_2 = 2$, it is squared exponential covariogram.

(2) Spherical Covariogram

$$C(\boldsymbol{h}) = \begin{cases} \theta_1(1 - 1.5(\|\boldsymbol{h}\|/\theta_2) + 0.5(\|\boldsymbol{h}\|/\theta_2)^3) & \text{if } \|\boldsymbol{h}\| < \theta_2 \\ 0 & \text{if } \|\boldsymbol{h}\| \geq \theta_2 \end{cases} \tag{1.3}$$

$\theta_1 > 0$ is the variance and θ_2 is the range. It is not a valid isotropic covariogram when $d \geq 4$. Spherical covairance function is only once differentiable.

(3) Matern Covariogram

$$C(\boldsymbol{h}) = \frac{\theta_1(\|\boldsymbol{h}\|/\theta_2)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\|\boldsymbol{h}\|/\theta_2) \tag{1.4}$$

where K_ν is the modified Bessel function of the second kind of order $\nu > 0$. θ_1 is the variance, θ_2 is the range parameter, ν is smoothness parameter. The Matern family does not have a closed form, and there are three special cases:

a. When $\nu = 0.5$, $C(\boldsymbol{h}) = \theta_1 exp(-\|\boldsymbol{h}\|/\theta_2)$. It is called *exponential covariogram*. The range parameter θ_2 has a practical interpretation: $3\theta_2$ is the practical range of the covariogram, at which the correlation is approximately 0.05.

b. When $\nu = 1.5$, $C(\boldsymbol{h}) = \theta_1(1 + \|\boldsymbol{h}\|/\theta_2)exp(-\|\boldsymbol{h}\|/\theta_2)$.

c. When $\nu \to \infty$, $C(\boldsymbol{h}) = \theta_1 exp(-\|\boldsymbol{h}/\theta_2\|^2)$. It is called *Gaussian covariogram*.

1.2.4 Kriging

Kriging is an interpolation method for spatial data, which is governed by the correlation structure. The result of kriging is the expected value and variance for every point within a region. In geostatistics, kriging is the best linear unbiased predictor (BLUP) based on observational data with weighted spatial covariance. It is named after Danie G. Krige ([Krige, 1951]) who first applied this technique to estimate the most likely distribution of gold in South Africa.

Suppose we observed some variable $\boldsymbol{Y} = (\boldsymbol{y}(\boldsymbol{s}_1), ..., \boldsymbol{y}(\boldsymbol{s}_n))'$ at n locations $(\boldsymbol{s}_1, ..., \boldsymbol{s}_n)$. To estimate the mean and variance of $\boldsymbol{y}(\boldsymbol{s}_0)$ at an unknown location $\boldsymbol{s}_0$, we consider the following linear predictors:

$$\hat{\boldsymbol{y}}(\boldsymbol{s}_0) = \boldsymbol{y}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i \boldsymbol{y}(\boldsymbol{s}_i) \tag{1.5}$$

where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)'$. To determine the weights $\boldsymbol{\lambda}$, we need assumptions on $\boldsymbol{y}$ and minimize the mean squared error (MSE), $E((\hat{\boldsymbol{y}}(\boldsymbol{s}_0) - \boldsymbol{y}(\boldsymbol{s}_0))^2)$ under the constraint

$E(\hat{\boldsymbol{y}}(\boldsymbol{s}_0)) = E(\boldsymbol{y}(\boldsymbol{s}_0))$. These assumptions distinguish among simple, ordinary and universal kriging. Assuming the mean function is $\mu(\boldsymbol{s}) = E(\boldsymbol{y}(\boldsymbol{s}))$ and deterministic, we will discuss various kriging methods.

(1) *Simple kriging* is kriging with known means. Assuming $E(\boldsymbol{y}(\boldsymbol{s})) = m_0$, the prediction at s_0 is

$$\hat{\boldsymbol{y}}(\boldsymbol{s}_0) = m_0 + K'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \mu) \tag{1.6}$$

The prediction variance or kriging variance, which is also the MSE, is:

$$\sigma^2_{sk}(\boldsymbol{s}_0) = Var(\boldsymbol{y}(\boldsymbol{s}_0)) - K'\boldsymbol{V}^{-1}K \tag{1.7}$$

where $K = Cov(\boldsymbol{y}(\boldsymbol{s}_0), \boldsymbol{Y})$ and $V = Var(\boldsymbol{Y})$.

(2) *Ordinary kriging* is kriging with unknown means. If the mean function at all locations is unknown but constant, we need to estimate the mean first, then apply *simple kriging* to get the covariance. Assuming $E(\boldsymbol{y}(\boldsymbol{s}_0)) = m$ where m is some unknown constant, the ordinary kriging(BLUP) and prediction variance at s_0 are:

$$\begin{aligned}
\hat{\boldsymbol{y}}(\boldsymbol{s}_0) &= \hat{\mu} + K'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \mathbf{1}\hat{\mu}) && (1.8)\\
\hat{\mu} &= \frac{\mathbf{1}'\boldsymbol{V}^{-1}\boldsymbol{Y}}{\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1}} \\
\sigma^2_{ok}(\boldsymbol{s}_0) &= Var(\boldsymbol{y}(\boldsymbol{s}_0)) - K'\boldsymbol{V}^{-1}K + \frac{(1 - \mathbf{1}'\boldsymbol{V}^{-1}K)^2}{\mathbf{1}'\boldsymbol{V}^{-1}\mathbf{1}} && (1.9)
\end{aligned}$$

The second term of $\sigma^2_{ok}(\boldsymbol{s}_0)$ is always nonnegative, so ordinary kriging has a larger variance than simple kriging. The reason is because the ordinary kriging needs to estimate the mean function. If we replace m_0 in the simple kriging predictor with BLUP, we get the ordinary kriging predictor.

(3) *Universal kriging* is kriging with a trend. If the mean function can be expressed as a linear function of some explanatory variables, it is called *universal kriging*.

Ordinal kriging is a special case of universal kriging. Suppose $\boldsymbol{X}$ is the matrix of explanatory variables at n locations, universal kriging will be:

$$\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{\beta} + \epsilon \tag{1.10}$$

Where $\boldsymbol{\beta}$ is the parameter to be estimated, and ϵ is the error term and captures the correlation structure in $\boldsymbol{Y}$. The resulting BLUP and prediction variance are:

$$\begin{aligned} \hat{\boldsymbol{y}}(\boldsymbol{s}_0) &= \boldsymbol{X}(\boldsymbol{s}_0)'\hat{\boldsymbol{\beta}} + K'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}'\hat{\boldsymbol{\beta}}) && (1.11) \\ \sigma^2_{uk}(\boldsymbol{s}_0) &= Var(\boldsymbol{y}(\boldsymbol{s}_0)) - K'\boldsymbol{V}^{-1}K && (1.12) \\ &+ (\boldsymbol{X}(\boldsymbol{s}_0) - \boldsymbol{X}\boldsymbol{V}^{-1}K)'(\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}')^{-1}(\boldsymbol{X}(\boldsymbol{s}_0) - \boldsymbol{X}\boldsymbol{V}^{-1}K) \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{Y}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. Comparing the variance structure of simple kriging and universal kriging, universal kriging has larger variance.

1.2.5 CoKriging

If we observe p variables and $p > 1$, there might be cross correlation between different variables in addition to spatial correlation. Then, we could model the correlation function properly and make predictions using other variables. This process is called *cokriging*.

Let $y_i(\boldsymbol{s})$ denote the i^{th} underlying spatial process at different locations, we want to predict y_1 at unknown location s_0 based on all observations $y_i(\boldsymbol{s}_j), i = 1, \cdots, p, j = 1, \cdots, n_i$. Then, y_1 is called the *primary variable* and the rest *auxiliary variables*. Let $\boldsymbol{y} = (y_1, y_2, \cdots, y_p)'$ and y_i is a vector of observations of the i^{th} variable, $K = Cov(\boldsymbol{y}, y_1(\boldsymbol{s}))$ and $\Sigma = Var(\boldsymbol{y})$. The cokriging is

$$\hat{y}_1(\boldsymbol{s}_0) = E(y_1(\boldsymbol{s}_0)) + K'\Sigma^{-1}(\boldsymbol{y} - E(\boldsymbol{y})) \tag{1.13}$$

With different assumptions on $E(y_1(\boldsymbol{s}_0))$, we have the following two types of *cokriging*.

(1) *Simple cokriging* is cokriging with known means

$$\hat{y}_1(\boldsymbol{s}_0) = E(y_1(\boldsymbol{s}_0)) + K'\Sigma^{-1}(\boldsymbol{y} - E(\boldsymbol{y})) \tag{1.14}$$

and variance

$$\sigma^2_{sck} = Var(y_1(\boldsymbol{s}_0)) - K'\Sigma^{-1}K$$

In general, simple cokriging variance is less than kriging variance with exception of proportional model ([Zhang, 2014]).

(2) *Ordinary cokriging* is cokriging with unknown constant mean.

Let $e_1 = (1, 0, \cdots, 0)_p$, the ordinary cokriging is

$$\hat{y}(\boldsymbol{s}_0) = e_1'(\boldsymbol{J}\Sigma^{-1}\boldsymbol{J}')^{-1}\boldsymbol{J}\Sigma^{-1}\boldsymbol{y} + K'\Sigma^{-1}\boldsymbol{y} - K'\Sigma^{-1}\boldsymbol{J}'(\boldsymbol{J}'\Sigma^{-1}\boldsymbol{J}')^{-1}\boldsymbol{J}\Sigma^{-1}\boldsymbol{y} \tag{1.15}$$

Let $\hat{\mu} = (\boldsymbol{J}\Sigma^{-1}\boldsymbol{J}')^{-1}\boldsymbol{J}\Sigma^{-1}\boldsymbol{y}$, the ordinary cokriging can be simplified as

$$\hat{y}(\boldsymbol{s}_0) = e_1'\hat{\mu} + K'\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{J}'\hat{\mu})$$

The prediction variance is

$$\sigma^2_{ock} = Var(Y_1(\boldsymbol{s}_0)) - K'\Sigma^{-1}K + (e_1 - \boldsymbol{J}\Sigma^{-1}K)(\boldsymbol{J}\Sigma^{-1}\boldsymbol{J}'(e_1 - \boldsymbol{J}\Sigma^{-1}K))$$

Comparing the prediction variance between simple cokriging and ordinary cokriging, the latter one has larger variance.

2. UNCERTAINTY QUANTIFICATION: SCALE VARIATION

2.1 Models for Different Scales

In ecosystem models, the input variables and output variables are usually time series at different temporal scales. When we are developing models, the prediction and data set in coarse scales are usually aggregated over finer scales. Then, which model is better in terms of predictability and accuracy as well as computational cost? Which temporal scale should we choose? To provide guidance for future research, we will compare the prediction accuracy and uncertainties from different temporal scale models. To illustrate the problem, we start with multiple linear regression models. Section 2.1.1 will discuss the scaling issues, and Section 2.1.2 will compare different temporal scale models theoretically, Section 2.2 will compare prediction of GPP from different temporal scale models.

2.1.1 The Scaling Issues

Suppose Y is the response variable to be regressed on $p-1$ explanatory variables $x_1, \ldots, x_{p-1}$. Each of the variables is observed at time points $t = 1, \ldots, n$, say daily. The linear regression model becomes

$$y_t = \beta_0 + \sum_{i=1}^{p-1} x_{t,i}\beta_i + \epsilon_t, t = 1, \ldots, n, \tag{2.1}$$

where the error terms ϵ_t are assumed to be i.i.d. $N(0, \sigma^2)$.

However, there are situations when the model is applied at a larger scale, say, weekly. The aggregated variables $y_t^{(w)} = \sum_{i=1}^{s} y_{s(t-1)+i}, x_{t,k}^{(w)} = \sum_{i=1}^{s} x_{s(t-1)+i,k}, k =$

$1, \ldots, p-1$ are used in the regression, where s denotes the time units the variables are aggregated upon (e.g., $s = 7$ for the weekly scale). The model becomes

$$y_t^{(w)} = s\beta_0 + \sum_{i=1}^{p-1} x_{t,i}^{(w)} \beta_i + \epsilon_t^{(w)}, t = 1, \ldots, m, \tag{2.2}$$

The two models share the same linear parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_{p-1}]'$, but the error terms in (2.2) has a larger variance than (2.1). In addition, there are fewer observations for the larger scale model (2.2). Hereafter, we assume that $n = ms$.

The two central questions this work concerned of are as followed. First, how do the two scales affect the estimation of the parameters β_i and the variance σ^2? Second, how do the scales affect the prediction? More specifically, suppose we want to predict $y_{m+1}^{(w)}$, we can obtain this prediction from both models. How different would these two predictions be?

2.1.2 Theoretical Results

In this section, we provide some theoretical results that allow us to draw some explicit conclusions. Denote by $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ the least squares estimators of $\boldsymbol{\beta}$ and σ^2, respectively, which are obtained by fitting model (2.1), and by $\hat{\boldsymbol{\beta}}^{(w)}$ and $\hat{\sigma}^{2(w)}$ the least squares estimators according to model (2.2). If we denote by $\boldsymbol{X}$ the design matrix in model (2.1) and by $\boldsymbol{y}$ the vector of response variable, then

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{(-1)}\boldsymbol{X}'\boldsymbol{y},\ \hat{\sigma}^2 = ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2/(n-p).$$

The design matrix $\boldsymbol{X}^{(w)}$ and the vector aggregated response variable $\boldsymbol{y}^{(w)}$ are related to $\boldsymbol{X}$ and $\boldsymbol{y}$ in the following way

$$\boldsymbol{X}^{(w)} = \boldsymbol{J}\boldsymbol{X},\ \boldsymbol{y}^{(w)} = \boldsymbol{J}\boldsymbol{y},$$

where $\boldsymbol{J} = I_m \otimes 1_s$ is an $m \times n$ matrix where I_m is an $m \times m$ identity matrix and 1_s is an s-dimension vector of all 1s. The estimates from model (2.2) can be written

$$\hat{\boldsymbol{\beta}}^{(w)} = (\boldsymbol{X}^{(w)'}\boldsymbol{X}^{(w)})^{(-1)}\boldsymbol{X}^{(w)'}\boldsymbol{y}^{(w)},\ \hat{\sigma}^{2(w)} = s||\boldsymbol{y}^{(w)} - \boldsymbol{X}^{(w)}\hat{\boldsymbol{\beta}}^{(w)}||^2/(m-p)$$

where s is the period of time units the large scale is aggregated upon.

The following proposition says that the smaller scale model yields more efficient estimators than the larger scale model.

Proposition 2.1.1 *Observing $y_1, \ldots, y_n$ with $n = ms$, the estimators given through the vector of response variable of two models (2.1) and (2.2) have the following properties.*

(i) Both $\hat{\boldsymbol{\beta}}^{(w)}$ and $\hat{\boldsymbol{\beta}}$ are unbiased estimators of $\boldsymbol{\beta}$ but the former is more efficient, i.e.,

$$E(\hat{\boldsymbol{\beta}}^{(w)}) = E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta},$$

and $Var(\hat{\boldsymbol{\beta}}^{(w)}) - Var(\hat{\boldsymbol{\beta}})$ is positive semi-definite.

(ii) Both $\hat{\sigma}^2$ and $\hat{\sigma}^{2(w)}$ are unbiased estimators of σ^2. In addition,

$$Var(\hat{\sigma}^2) = \frac{2\sigma^4}{n-p},\ Var(\hat{\sigma}^{2(w)}) = \frac{2\sigma^4}{m-p}.$$

Hence $\hat{\sigma}^2$ is more efficient.

The Proposition readily follows the Gauss-Markov theorem [Stapleton, 1995]. We only sketch the proof here. It is obvious that both $\hat{\boldsymbol{\beta}}^{(w)}$ and $\hat{\boldsymbol{\beta}}$ are unbiased. Since $\hat{\boldsymbol{\beta}}^{(w)}$ is a linear unbiased estimator, the Gauss-Markov theorem implies that $\hat{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}^{(w)}$. It is well known that $||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2/\sigma^2$ has a χ^2-distribution with $(n-p)$ degrees of freedom. It follows

$$Var(\hat{\sigma}^2) = \frac{2\sigma^4}{n-p}.$$

Indeed, the above can be found in classical textbooks on regression. Similarly, because

$$\frac{||\boldsymbol{y}^{(w)} - \boldsymbol{X}^{(w)}\hat{\boldsymbol{\beta}}^{(w)}||^2}{\sigma^4/s}$$

has a χ^2-distribution with $(m-p)$ degrees of freedom with a variance $2(m-p)$, it follows that

$$Var(\hat{\sigma}^{2(w)}) = \frac{2\sigma^4}{m-p}.$$

Next, we consider the effects of scales on prediction. If we observe the explanatory variables at s consecutive time points, $n+1, \ldots, n+s$, and want to make a prediction of the aggregated response variable $y^{(w)}$, we could obtain the prediction in two ways, using the two models (2.1) and (2.2). The explanatory variables for the larger scale model is $\boldsymbol{x}_{m+1}^{(w)} = \sum_{t=n+1}^{n+s} \boldsymbol{x}_i$, where $\boldsymbol{x}_i$ is the vector of explanatory variables at time i. We could get the prediction of y from the two different temporal scale models as follows:

$$\hat{Y} = \sum_{i=n+1}^{n+s} \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} = \boldsymbol{x}^{(w)'} \hat{\boldsymbol{\beta}}. \tag{2.3}$$

$$\hat{Y}^{(w)} = \boldsymbol{x}^{(w)'} \hat{\boldsymbol{\beta}}^{(w)}. \tag{2.4}$$

Comparison of the two predictions is given in the following proposition.

Proposition 2.1.2 *Under the formulation of models (2.1) and (2.2), the two predictors (2.3) and (2.4) have the following properties:*

(i) $E(\hat{Y}) = E(\hat{Y}^{(w)}) = \boldsymbol{x}^{(w)'} \boldsymbol{\beta}$.

(ii) $Var(\hat{Y}^{(w)}) \geq Var(\hat{Y})$.

This proposition follows from the unbiasedness of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{(w)}$, and the fact that $\hat{\boldsymbol{\beta}}$ is the best unbiased linear estimator of $\boldsymbol{\beta}$. Indeed, the Gauss-Markov theorem implies that for any vector $\boldsymbol{x}$,

$$Var(\boldsymbol{x}' \hat{\boldsymbol{\beta}}^{(w)}) \geq Var(\boldsymbol{x}' \hat{\boldsymbol{\beta}}).$$

2.1.3 Scaling Issues with Polynomial Regression

In this section, we consider the scaling issue in the polynomial regression. What complicates in this case is that there are two possible ways to run the model at the larger scale. Suppose the regression model at the smaller scale is

$$y_t = \beta_0 + \sum_{i=1}^{p-1} x_{t,i} \beta_i + \sum_{(i,j) \in \Delta} x_{t,i} x_{t,j} \beta_{ij} + \epsilon_t, t = 1, \ldots, n, \tag{2.5}$$

where Δ is an index set for the high order term. For example, $\Delta = \{(i,j), i,j = 1,\ldots,p-1, i \neq j\}$ if all second order terms are included in the model.

One way to formulate the larger scale model is to aggregate all variables as in model (2.2)

$$y_t^{(w)} = s\beta_0 + \sum_{i=1}^{p-1} x_{t,i}^{(w)}\beta_i + \sum_{(i,j)\in\Delta} x_{t,ij}^{(w)}\beta_{ij} + \epsilon_t^{(w)}, t = 1,\ldots,m, \tag{2.6}$$

where $y_t^{(w)}$ and $x_{t,i}^{(w)}$ are defined the same as in (2.2), $x_{t,ij}^{(w)} = \sum_{k=1}^{s} x_{s(t-1)+k,i}\, x_{s(t-1)+k,j}$ is the aggregated cross product $x_{t,i}x_{t,j}$. Comparison between models (2.5) and (2.6) follows the discussion in the previous section. We can say that model (2.5) at the smaller scale results in more efficient estimation and better prediction.

In practice, however, the larger scale model is often run as follows.

$$y_t^{(w)} = \beta_0^{(w)} + \sum_{i=1}^{p-1} x_{t,i}^{(w)}\beta_i^{(w)} + \sum_{(i,j)\in\Delta} x_{t,i}^{(w)}x_{t,j}^{(w)}\beta_{ij}^{(w)} + \epsilon_t^{(w)}, t = 1,\ldots,m, \tag{2.7}$$

where $x_{t,i}^{(w)}$ is same as defined previously. The high order terms are now aggregated differently. The larger scale model (2.7) and the small scale model (2.5) have different sets of parameters. Therefore, unlike in the previous section, a direct comparison between the two models is difficult if not impossible. For example, it does not make sense to compare the efficiency of estimators because the parameters in the two models are different. Similarly, for prediction, the two models assume different expected value to start with. Therefore, the two models may yield different prediction results.

The example in the next section reveals that the predicted value given by the larger scale model may be either smaller or larger than that given by the smaller scale model.

2.2 An Example

In this section, we consider an example of real data set, which motivated this work and also helps to show the difference the scales can make to statistical inferences. The response variable in this example is the gross primary production (GPP), which is the

total amount of energy primarily produced by plants through photosynthesis. The GPP can be calculated from the observations at the eddy flux towers. However, for a region such as a country or continent, the GPP has to be estimated by employing either statistical models or ecosystem models, which may range in complexity from empirical models (e.g., [Xiao et al., 2010], [Yang et al., 2007]) to biogeochemical models (e.g.,[Prince and Goward, 1995], [Running et al., 2004], [Turner et al., 2004]). Linear regression models have been employed to estimate the regional GPP. For example, [Zhang et al., 2007] used an empirical piecewise regression model to map GPP for the Northern Great Plains grasslands from flux tower measurements. [Xiao et al., 2010] developed an upscaling model based on the regression tree method to extrapolate eddy flux GPP data to the continental scale and producing continuous GPP estimates across multiple biomes. [Mueller et al., 2010] studied the variability of carbon flux measurement across different temporal scales. We will examine estimations of regional GPP given by models of different time scales.

2.2.1 Data and Model

We used the data collected at the AmeriFlux towers at 70 sites (`http://ameriflux.ornl.gov/`). We obtained the level 4 data from `http://cdiac.ornl.gov/ftp/ameriflux/data/Level4/`. The data consist of observations collected every half hour ranging from 2000 to 2007 at each site. The response variable is GPP and six explanatory variables are air temperature, global radiation, precipitation, vapor pressure deficit, land-cover type and enhanced vegetation index (EVI). These six variables were chosen based on previous studies. The first five variables were observed at the AmeriFlux sites and EVI was calculated from the Moderate Resolution Imaging Spectroradiometer (MODIS) every 8 days, which is the reason we choose the 8-day scale instead of the weekly scale. The land-cover type is a qualitative variable with 6 levels representing 6 land-cover categories. Based on these data, we fitted a poly-

Table 2.1.: The predicted annual GPP(Units: $TgCyr^{-1}$) over the US by year.

Year	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
Daily	6328	6109	6528	6587	6697	6299	6673
8-Day	6338	6133	6572	6604	6753	6348	6728
Monthly	5770	5509	5903	5941	6128	5744	6084

nomial regression of order 2 from (2.7) at three different scales: daily, 8-day, and monthly. We therefore have three fitted regression models.

To predict GPP at a site that is not part of AmeriFlux net, we use data from the North American Regional Reanalysis (NARR) (`http://www.emc.ncep.noaa.gov/mmb/rreanl/`). This data set has a spatial resolution of 0.5×0.5 degrees over the conterminous US, and the time range is 2001-2007. In total, the whole US has 3252 pixels. We predict the GPP at each of the pixel using the three fitted models and calculated the total GPP over the US by adding the pixel-level GPP.

2.2.2 Results

The first conclusion we can draw is that a large scale model can result in larger or smaller prediction. This can be seen in Table 2.1 which summarized the total GPP over the US for each year. We see that the 8-day model yields higher total GPP than the daily model in each of the seven years while the monthly model yields lower total GPP than the daily model. Figure 2.1 shows three predicted monthly total GPP over the US for each month between 2001 and 2007 in the whole US, from which we can see that the predicted monthly GPP from the three different temporal scale models are different. The 8-day model consistently provides higher predicted total GPP, which is consistent to what we observed from Table 2.1.

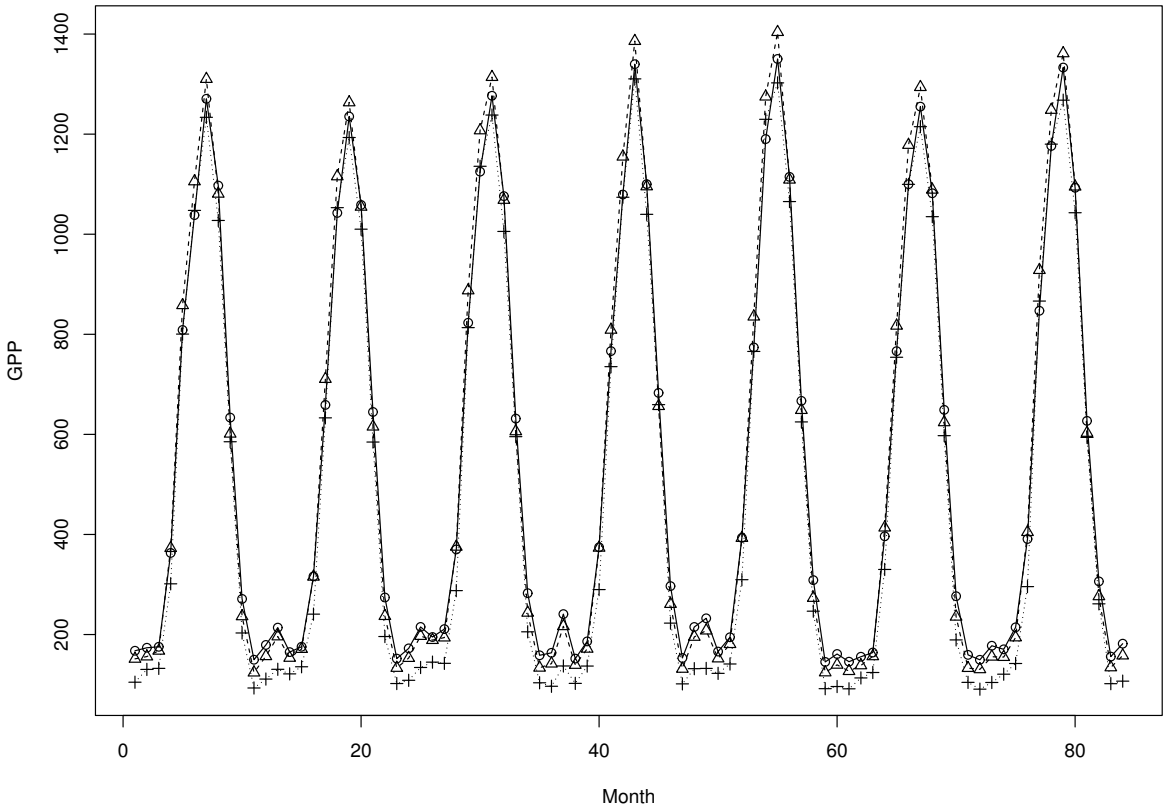

Figure 2.1.: Predicted monthly GPP (TgC) across 2001-2007 given by the daily model ($\circ$), the 8-day model ($\triangle$), and the monthly model ($+$).

In Figure (2.2), we plot the predicted annual GPP for the year 2007 at each pixel. The three different models reveal about the same spatial trend, but a careful examination also reveals some differences of the predicted GPPs in some areas.

Next we compare the prediction variances given by the three models at each pixel. Figure 2.3 plots the standard errors given by the three different temporal scale models at each pixel for year 2007. It is evident that the prediction error is smaller for finer resolutions, although we cannot justify this theoretically in this case.

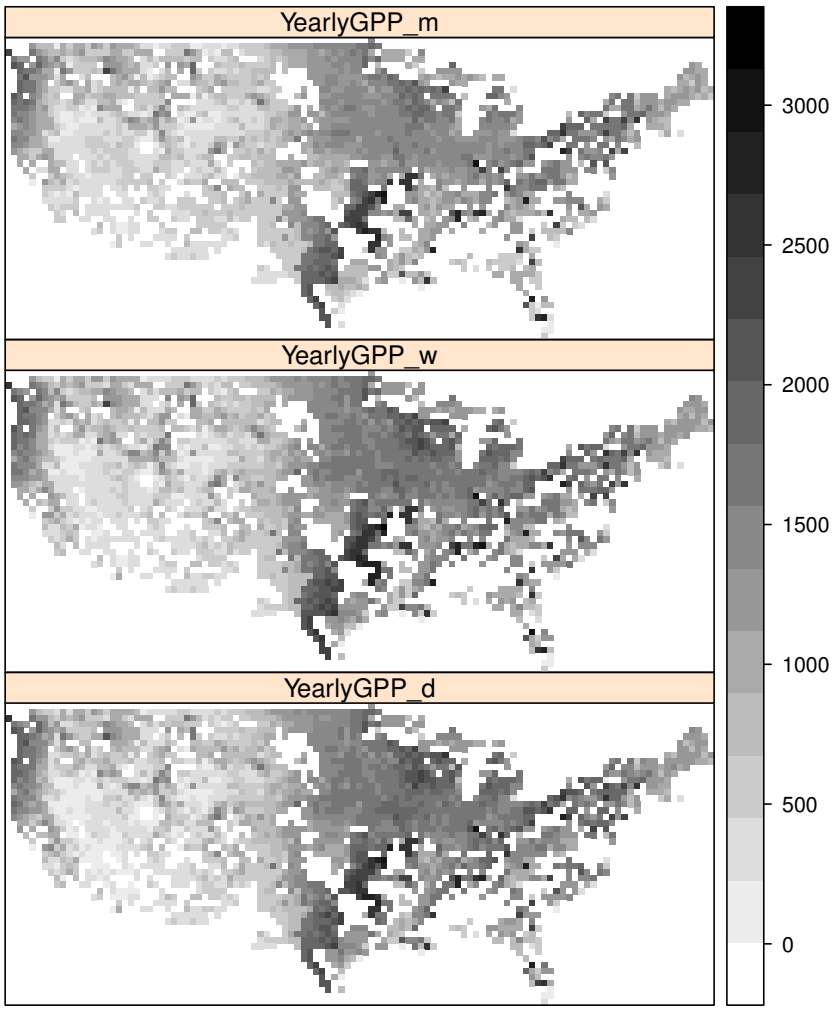

Figure 2.2.: Annual GPP(Units: $gCm^{-2}yr^{-1}$) predicted by three models for year 2007: monthly model (top), 8-day model (middle) and daily model (bottom).

2.3 Discussion

In this chapter, we provided some theoretical discussions on the scale issue in linear models. When there is no high order terms in the model, the smaller scale model is preferred whenever possible. However, if the model includes high order terms of the explanatory variables, direct comparisons are difficult and no explicit conclusions are given in this paper. The example revealed that a larger scale model

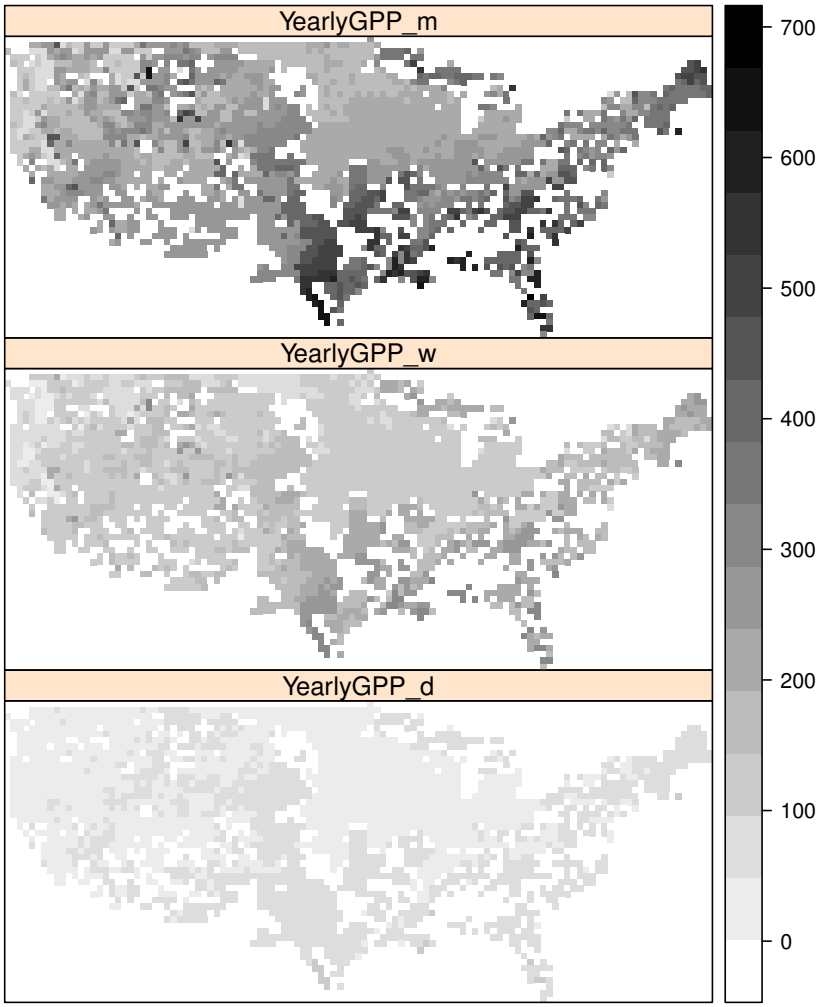

Figure 2.3.: Standard error of of GPP(Units: $gCm^{-2}yr^{-1}$) at each pixel for year 2007: monthly model (top), 8-day model (middle) and daily model (bottom).

can yield either larger or smaller predictions. For the polynomial regression, it would be an interesting problem to provide some conditions under which the larger scale model yields larger predictions, or conditions under which the larger scale model yields smaller predictions. It would be also interesting to investigate how the scales affect the prediction variance.

The prediction of GPP does not vary significantly across different temporal scales, but the uncertainties are different significantly. In more complexed models than regression models, we should choose an appropriate temporal scale balancing accuracy and computational cost. The TEM is at monthly temporal scale. In the following chapters, we develop our models at monthly temporal scale.

3. MULTI-OUTPUT EMULATOR FOR TERRESTRIAL ECOSYSTEM MODEL

3.1 Review of Methods

A *physical model* refers to a mathematical model that is built partially based upon physical or biological principles. It is a deterministic model and is usually defined by a set of differential equations. Physical models are used widely in atmospheric sciences, hydrology, ecology, and ecosystem studies. Because a physical model is deterministic, quantification of its uncertainty is of vital interest. The uncertainties may be attributed to inaccurate input data, insufficient knowledge of the parameters of the model, inherent randomness of the physical system, simplification of the model structure and so on.

Let $\boldsymbol{y} = \boldsymbol{u}(t, \boldsymbol{x}, \boldsymbol{\theta})$ be the deterministic output of a physical model where t and $\boldsymbol{x} \in R^2$ represent time and input variables, respectively, and $\boldsymbol{\theta}$ is the model parameters. One source of uncertainty is the uncertainty in the parameters $\boldsymbol{\theta}$ which is usually unknown and needs to be estimated in some way. Therefore the parameters can be regarded as random variables whose distribution G is either assumed to be known or unknown. The study of how the distribution of $Y(t, \boldsymbol{x}) = \boldsymbol{u}(t, \boldsymbol{x}, \boldsymbol{\vartheta})$ depends on G is the essential part of the uncertainty analysis. $\boldsymbol{\theta}$ is used to denote the random variables, and $\boldsymbol{\vartheta}$ is a particular value of $\boldsymbol{\theta}$. The function $\boldsymbol{u}$ usually is determined in a complex way such as by differential equations. The structure of $\boldsymbol{u}$, or equivalently, the structure of the differential equations, is another source of uncertainty for the model output.

To quantify the uncertainty in model structure is a difficult problem and has not been dealt with directly. However, it is possible to combine information from ensembles of multiple models as a way to indirectly assess the uncertainty in the

model. It borrows strength from different models. [Tebaldi et al., 2005] developed a Bayesian framework to combine a multi-model ensemble with observations to quantify uncertainty. I will review the methods to quantify the uncertainties attributed to the the parameters.

(1) Ensemble

If the model can be run thousands of times in a reasonable time, it is sometimes called a cheap model. For a cheap model, simulation is the major technique to quantify the uncertainty. Basically, the samples are simulated from the parameter distribution G. For each sample value of $\boldsymbol{\theta}$, the model is run to yield the output $\boldsymbol{y}(t, \boldsymbol{x}) = \boldsymbol{u}(t, \boldsymbol{x}, \boldsymbol{\theta})$. The outputs so obtained are also called an *ensemble*. This *ensemble* approach is usually the first choice of quantification of uncertainty if the computation is affordable.

The probability distribution of the ensemble reveals the variation and therefore the uncertainty of the model output. One may further study the contribution of each individual parameter $\boldsymbol{\theta}_i$ to the uncertainty, which is usually called the *sensitivity analysis*.

(2) Emulation

However, there are many cases where the model is so complex that the calculation of the output $\boldsymbol{y}(t, \boldsymbol{x})$ for any given set of input values is computationally heavy and it is precluded to simulate a large number of model outputs. In this case, suppose one can afford to obtain output at n design points $(t_1, \boldsymbol{x}_1, \boldsymbol{\vartheta}_1), ..., (t_n, \boldsymbol{x}_n, \boldsymbol{\vartheta}_n)$. One would use this limited amount of data to quantify the uncertainty of $\boldsymbol{u}(t, \boldsymbol{x}, \boldsymbol{\theta})$ for any given $(t, \boldsymbol{x}, \boldsymbol{\theta})$. Obviously, this is possible only if some statistical relationship between $\boldsymbol{u}(t, \boldsymbol{x}, \boldsymbol{\theta})$ and $\boldsymbol{u}(t_i, \boldsymbol{x}_i, \boldsymbol{\vartheta}_i)$ is postulated. Statistical *emulation* is a technique for this purpose. In particular, Gaussian process emulation has been used in the Bayesian

approach to quantifying uncertainty(e.g., [Oakley and O'Hagan, 2004] and [Conti and O'Hagan, 2010]).

In the Gaussian process emulation, it is assumed that $\boldsymbol{\theta}$ is Gaussian and, given $\boldsymbol{\theta}$, $Y(t, \boldsymbol{x})$ is a Gaussian process with a parametric mean function and some covariance function, which depend on two separate sets of parameters. The mean function is usually assumed to be a linear function $h(t, \boldsymbol{x}, \boldsymbol{\theta})'\beta$ for some parameter β and the covariance function may be stationary, i.e., depending only on the Euclidean distance. When the output is obtained at n design points $(t_1, \boldsymbol{x}_1, \boldsymbol{\vartheta}_1), ..., (t_n, \boldsymbol{x}_n, \boldsymbol{\vartheta}_n)$, the posterior mean $E(u(t, \boldsymbol{x}, \boldsymbol{\theta})|Y)$ and variance $V(u(t, \boldsymbol{x}, \boldsymbol{\theta})|Y)$ can be calculated, both of which quantify the uncertainty of the output $u(t, \boldsymbol{x}, \boldsymbol{\theta})$, where Y represents the random output at the n design points. The spatial linear interpolation method, *kriging*, may be applied, which would require that the covariance functions be known. To account for uncertainty in the mean and covariance function, [Oakley and O'Hagan, 2004] and [Conti and O'Hagan, 2010] considered a Bayesian approach and studied an efficient computational method for calculating the posterior distributions.

There are two choices that have to be made in the statistical emulation. One is the choice of the design points and the other is the choice of the spatial covariance function. In practice, the design points should be spread to cover the input space. [Sacks et al., 1989] discussed the choice of design points. [Conti and O'Hagan, 2010] used Latin Hypercube Sampling (LHS) method to get the design points.The Gaussian covariance function has been applied in many applications. However, it is too smooth and is known to have undesirable properties [Stein, 1999]. The Matern family with an adjustable parameter for the smoothness of the process is a better choice.

(3) Polynomial Chaos

Another approach to the quantification of uncertainty is centered around the approximation to stochastic differential equations that decouples the random part and

the deterministic part of the model output. Suppose the model is given by the differential equations

$$\frac{\partial u}{\partial t}(t, \boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{L}(u)$$

With initial and boundary conditions, $\boldsymbol{\theta}$ represents model parameters and is regarded as random, t and $\boldsymbol{x}$ represent time and input variables, $\mathcal{L}$ is a general operator. Hence the model output $u(t, \boldsymbol{x}, \boldsymbol{\theta})$ is random and allows the following generalized polynomial chaos(gPC) representation

$$u(t, \boldsymbol{x}, \boldsymbol{\theta}) = \sum_k u_k(t, \boldsymbol{x}) \Phi_k(\boldsymbol{\theta}) \tag{3.1}$$

where $\Phi_k(\boldsymbol{\theta})$, $k = 1, 2, ...$ are polynomials of $\boldsymbol{\theta}$, orthogonal with each other in the sense that $E(\Phi_j(\boldsymbol{\theta})\Phi_k(\boldsymbol{\theta})) = \delta_{jk}$. The particular polynomials satisfying this orthogonality depend on the distribution of $\boldsymbol{\theta}$ (e.g., [Xiu and Em Karniadakis, 2002]). The coefficients are deterministic and are determined by the structure of the differential equations governing u. In practice, the polynomial chaos is truncated at some finite terms, resulting in an approximation to the solution $u(t, \boldsymbol{x}, \boldsymbol{\theta})$. Therefore, once the coefficients $u_k(t, \boldsymbol{x})$ are determined, Monte Carlo samples can be generated efficiently through the approximation.

There are primarily two ways to determine the coefficients in (3.1): the Galerkin projection method and the collocation method. The former method usually yields better approximation results but involves solving a new deterministic differential equation while the latter does not need to solve new differential equations. However, the curse of dimensionality arises with the collocation method when the dimension of $\boldsymbol{\theta}$ increases.

3.2 Multi-Output Emulator

Emulator is an approximation of complex computer model, but faster to run than the computer model. It gives a probability distribution of the model output, and make predictions at training data points without uncertainty. For untried data points, it makes predictions and provides associated uncertainties. When the output is time series as $\boldsymbol{y}_{1:T}$, [Conti and O'Hagan, 2010] suggests the multi-output emulation method. This emulator is computationally cheap but statistically rigorous.

In Gaussian process emulation, the essential idea is to treat the model $\boldsymbol{y} = \boldsymbol{u}(\boldsymbol{x})$ as a black-box which is then modeled as a multivariate Gaussian process. This Gaussian process has a specified mean and covariance function which depends on two sets of parameters, where $\boldsymbol{y}$ is the output of the model and $\boldsymbol{x}$ is the input variables of the model. Given the model outputs $\boldsymbol{y}(\boldsymbol{x}_1), \ldots, \boldsymbol{y}(\boldsymbol{x}_n)$ at n design points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the model output at any input value $\boldsymbol{x}$ will be approximated by the conditional mean of $\boldsymbol{y}(\boldsymbol{x})$ given $\boldsymbol{y}(\boldsymbol{x}_1), \ldots, \boldsymbol{y}(\boldsymbol{x}_n)$. It is represented as follows

$$\boldsymbol{y}(\boldsymbol{x}) \sim GP(\mathbf{m}(\boldsymbol{x}), c(\boldsymbol{x}, \boldsymbol{x}')), \boldsymbol{y}(\boldsymbol{x}) \in \mathbb{R}^q \tag{3.2}$$

where the mean function $\mathbf{m}(\boldsymbol{x}) = \mathbf{B}^T h(\boldsymbol{x})$, $h(\boldsymbol{x}) \in \mathbb{R}^m$, and B is a $m \times q$ matrix. For example, $h(\boldsymbol{x}) = (1, x_1, x_2, ..., x_p)^T, m = p + 1$.

The covariance function $c(\boldsymbol{x}, \boldsymbol{x}')$ describes the correlation between $\boldsymbol{y}(\boldsymbol{x})$ and $\boldsymbol{y}(\boldsymbol{x}')$. The most commonly used multivariate correlation function are *proportional model, linear coregionalization model* and *multivariate Matern model*. Here, proportional model is chosen for computational efficiency [Conti and O'Hagan, 2010],

$$c(\boldsymbol{x}_1, \boldsymbol{x}_2) = \rho(\boldsymbol{x}_1, \boldsymbol{x}_2)V.$$

where V is a $q \times q$ positive-definite symmetric matrices, $\rho(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T R(\boldsymbol{x}_1 - \boldsymbol{x}_2))$. R is a diagonal matrix with diagonal elements $\boldsymbol{r} = (r_1, ..., r_p)$.

To get the training sample, the computer model is run at a pre-selected design set $S = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$ and denote the simulated output by matrix $D = [\boldsymbol{u}(\boldsymbol{s}_r)] \in \mathbb{R}_{n,q}$. The design points are sampled from the distributions of $\boldsymbol{x}$. The data D has matrix-normal

distribution $\mathcal{N}_{n,q}(HB, A, V)$ ([Rowe, 2002]) with mean HB and covariance matrix being the Kronecker product $A \otimes V$, where $H = [\mathbf{h}(\boldsymbol{s}_1), .., \mathbf{h}(\boldsymbol{s}_n)]$ and $A = [\rho(\boldsymbol{s}_l, \boldsymbol{s}_r)]_{n,n}$. It follows the property of normal distributions that $\boldsymbol{u}(\boldsymbol{x})$:

$$\boldsymbol{u}(\boldsymbol{x})|B, V, \boldsymbol{r}, D \sim GP(m^*(\boldsymbol{x}), \rho^*(\boldsymbol{x}_1, \boldsymbol{x}_2)V) \tag{3.3}$$

Here, for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{X}^p$

$$\begin{aligned} m^*(\boldsymbol{x}_1) &= B^T\mathbf{h}(\boldsymbol{x}_1) + (D - HB)^T A^{-1}\mathbf{t}(\boldsymbol{x}_1) \\ \rho^*(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \rho(\boldsymbol{x}_1, \boldsymbol{x}_2) - \mathbf{t}^T(\boldsymbol{x}_1)A^{-1}\mathbf{t}(\boldsymbol{x}_2) \end{aligned}$$

with $\mathbf{t}^T = [\rho(., \boldsymbol{s}_1), ..., \rho(., \boldsymbol{s}_n)]$.

Assuming the prior distribution of $(B, V, \boldsymbol{r})$:

$$p(B, V, \boldsymbol{r}) \propto \Pi_{i=1}^p \frac{1}{(1 + r_i)^2} |V|^{-(q+1)/2}. \tag{3.4}$$

B will be integrated out of (3.3) to get

$$\boldsymbol{u}(\boldsymbol{x})|V, \boldsymbol{r}, D \sim GP(m^{**}(\boldsymbol{x}), \rho^{**}(\boldsymbol{x}_1, \boldsymbol{x}_2)V) \tag{3.5}$$

Here, for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{X}^p$

$$m^{**}(\boldsymbol{x}_1) = \hat{B}_{GLS}^T\mathbf{h}(\boldsymbol{x}_1) + (D - H\hat{B}_{GLS}^T)^T A^{-1}\mathbf{t}(\boldsymbol{x}_1) \tag{3.6}$$

$$\begin{aligned} \rho^{**}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \rho^*(\boldsymbol{x}_1, \boldsymbol{x}_2) + [\mathbf{h}(\boldsymbol{x}_1) - H^T A^{-1}\mathbf{t}(\boldsymbol{x}_1)]^T \\ \cdot (H^T A^{-1} H)^{-1}[\mathbf{h}(\boldsymbol{x}_1) - H^T A^{-1}\mathbf{t}(\boldsymbol{x}_2)] \end{aligned} \tag{3.7}$$

where $\hat{B}_{GLS} = (H^T A^{-1} H)^{-1} H^T A^{-1} D$ is the generalized least squares(GLS) estimator of B. The conditional posterior distribution of $\boldsymbol{u}$ given $\boldsymbol{r}$ is the q-variate T process with $n - m$ degree of freedom:

$$\boldsymbol{u}(\boldsymbol{x})|\boldsymbol{r}, D \sim \mathcal{T}(m^{**}(\boldsymbol{x}), \rho^{**}(\boldsymbol{x}, \boldsymbol{x})\hat{V}_{GLS}; n - m) \tag{3.8}$$

where,$\hat{V}_{GLS} = (n - m)^{-1}(D - H\hat{B}_{GLS}^T)^T A^{-1}(D - H\hat{B}_{GLS}^T)$ is the GLS estimator of V.

The final step is to estimate the range parameters $\boldsymbol{r}$. Using Bayes' theorem, the posterior distribution of the parameters $(B, V, \boldsymbol{r})$ are:

$$\begin{aligned} p(B,V,\boldsymbol{r}|D) &\propto f(D|B,V,\boldsymbol{r})\cdot p(B,V,\boldsymbol{r}) \\ &\propto \Pi_{i=1}^{p}\frac{1}{(1+r_i)^2}\cdot|V|^{-\frac{q+1}{2}}\cdot|A|^{-q/2} \\ &\quad\cdot|\Sigma|^{-n+-\frac{q+1}{2}}\exp\{-\frac{1}{2}S_R^2\} \end{aligned} \tag{3.9}$$

Here, $S_R^2 = tr\{A^{-1}(D-HB)\Sigma^{-1}(D-HB)'\}$. After integrating out (B,V) from (3.9), the posterior distribution is

$$\begin{aligned} p(\boldsymbol{r}|D) &\propto \Pi_{i=1}^{p}\frac{1}{(1+r_i)^2}\cdot|A|^{-q/2} \\ &\quad\cdot|H^TA^{-1}H|^{-q/2}|D^TGD|^{-\frac{n-p-1}{2}} \end{aligned} \tag{3.10}$$

The mode or median of the posterior distribution (3.10) can be used to estimate $\boldsymbol{r}$.

3.3 Application: Terrestrial Ecosystem Model

3.3.1 Terrestrial Ecosystem Model

The Terrestrial Ecosystem Model is an ecosystem model which describes the dynamics of carbon, nitrogen, soil, water and other vegetation variables. The model has been developed and examined to describe the global carbon dynamics (e.g. [Zhuang et al., 2003]). It uses space time variables such as temperature, precipitation, cloudiness, to make monthly estimates of carbon and nitrogen fluxes. Other fixed input variables are CO2, vegetation type, location and latitude. The output variables most people are interested in is net primary production (NPP) which is the difference between gross primary production (GPP) and the vegetation respiration.

NPP is mostly modeled by process-based Terrestrial Ecosystem Model. The main random variables which relates to NPP are temperature, precipitation and global radiation. Here, I will consider emulating the monthly NPP($gm^{-2}month^{-1}$) using multi-output Gaussian process emulation with fixed parameters $\boldsymbol{\theta}$. The vegetation type considered is "temperate deciduous forest".

3.3.2 Data

Let NPP be denoted by $\boldsymbol{y} \in R^q$ where $q = 12$ represents 12 months. For a fixed location, the input variables $\boldsymbol{x}$ are temperature (Jan-Dec), precipitation (Jan-Dec) and global radiation (Jan-Dec). Other variables such as location, soil type, vegetation type are fixed.

In this study, the input variables $\boldsymbol{x}$ is assumed to follow multivariate normal distribution in which the mean and covariance are estimated from historical monthly data `http://hydrology.princeton.edu/data/pgf/0.5deg/monthly/`. To get the training data set, $n = 200$ data points are sampled from the multivariate normal distribution, which spread out the input space. Compared with the LHS method, our approach is much closer to the real environmental data for the same number of simulation runs.

The model output is obtained by running TEM for each training data point, and is denoted by D. Then, the method described in Section 3.2 is applied to compute the posterior distribution of NPP including posterior mean and covariance. Furthermore, $n_v = 200$ validation data are simulated to validate the statistical emulator. The spatial covariance function is a separable covariance function.

3.3.3 Results

Here, the output $\boldsymbol{y}(\boldsymbol{x})$ is monthly NPP in one year. To get data D, the first step is to sample training data from the input space. There are many ways to sample training data, like Latin Hypercube sampling method used in [Conti and O'Hagan, 2010]. Next, D is the output of computer model with the training samples. Then, to evaluate the performance of our emulator, $n_v = 200$ validation data points are sampled from the input space. In Fig. 3.1, I compared the emulated NPP with the validation sample. It shows that the emulated NPP is very close to TEM model output.

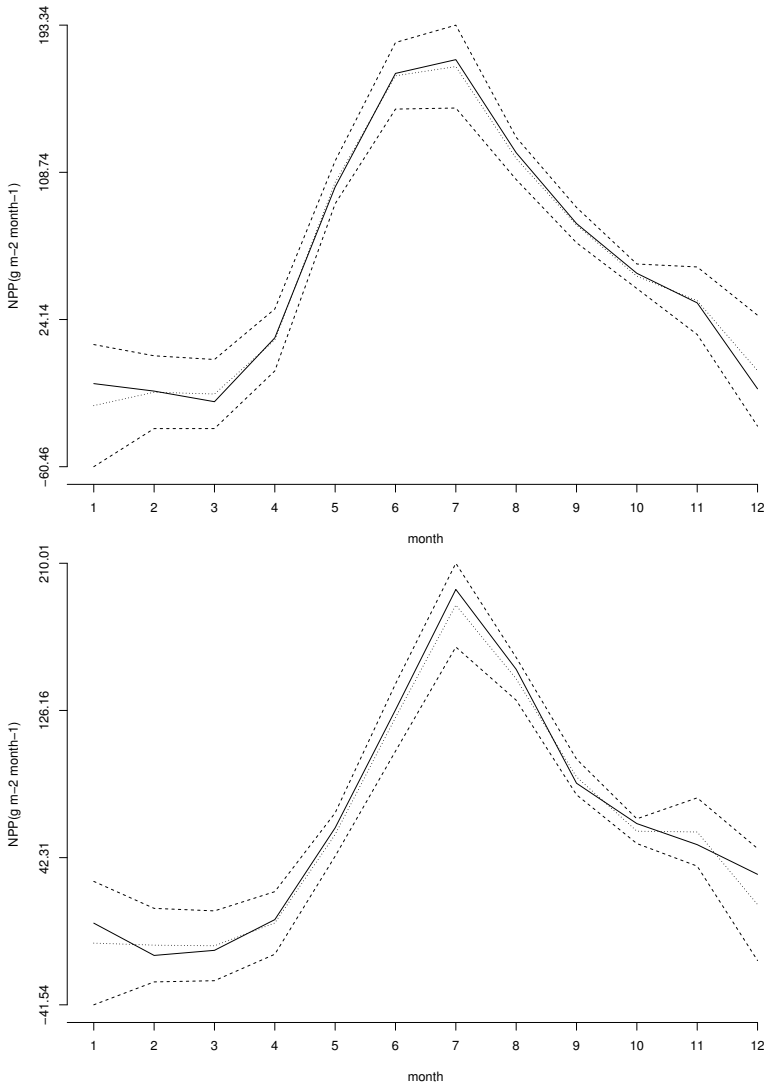

Figure 3.1.: Validation output(—) and emulated ($\cdots$) NPP. (- -) is the 95% confidence band.

3.3.4 Model Validation

To quantify how accurate the *emulator* approximating the TEM model output, the actual coverage rate of the 95% confidence interval are calculated, which is the percentage of NPP falling into the confidence intervals. The actual coverage rate is 95.5%, which is very close to the theoretical rate of 95%.

Table 3.1.: The root mean squared standardized error for each month

Jan	**Feb**	**Mar**	**April**	**May**	**Jun**
0.920	0.770	1.028	1.005	0.862	0.892
Jul	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
.941	0.977	1.044	0.931	0.978	0.999

The another measure to quantify the performance of statistical emulation is the root mean squared standardized error(RMSSE) defined below. For a validation $\boldsymbol{x}$, define the vector of standardized errors $s(\boldsymbol{x}_i) = (s(\boldsymbol{x}_i)_1, \cdots, s(\boldsymbol{x}_i)_q)^T \in \mathbb{R}^q$,

$$s(\boldsymbol{x}_i)_j = \frac{\boldsymbol{u}(\boldsymbol{x}_i)_j - \mathbf{m}^{**}(\boldsymbol{x}_i)_j}{\sqrt{\rho^{**}(\boldsymbol{x}_i, \boldsymbol{x}_i)\sigma_{jj},}}, j = 1, 2, \cdots, q \tag{3.11}$$

$$RMSSE_j = \sqrt{\frac{1}{n_v}\sum_{i=1}^{n_v}(s(\boldsymbol{x}_i)_j)^2}, j = 1, 2, \cdots, q \tag{3.12}$$

Table 3.1 shows us the twelve month's RMSSE. For a good model fit, the RMSSE should be close to 1 (e.g., [Conti and O'Hagan, 2010]). From Table 3.1, the emulator mimics the model output very well.

3.3.5 Summary

In this section, I first described a statistical approach for quantifying uncertainty of an ecosystem model through a Bayesian framework. The results show that the emulator approximates TEM output very closely and therefore reliably quantifies the uncertainty of the model. A separable multivariate covariance function is used following some existing work [Conti and O'Hagan, 2010, Oakley and O'Hagan, 2002, 2004]. Although this separability of covariance function simplifies computation significantly, it may not yield the best predictive performance. An interesting problem for future work is to use a more flexible and less restricted multivariate covariance function.

In Chapter 4 and Chapter 5, I will use Bayesian emulation approach to calibrate unknown parameters and quantify uncertainties in the physical model.

4. CALIBRATION OF PHYSICAL MODELS

4.1 Introduction

In physical models, there are two types of inputs: input variables (e.g. temperature, precipitation, cloudiness) and input parameters (e.g. maximum rate of photosynthesis at 0 °C). The latter one is assumed to be fixed and unknown. Usually, physical models are implemented as computer models (*simulators*) in which the parameters are chosen in advance. Since all models are mainly approximating the real physical process, there are discrepancies between a physical model and the real process. Calibration is the process of adjusting the unknown parameters to minimize this discrepancy, and methods have been developed to calibrate the parameters. For example, if y is defined as output from computer model, z as real observation and $\boldsymbol{\theta}$ as the unknown parameters, least squares calibration is to find $\boldsymbol{\theta}$ which minimizes the discrepancy

$$\hat{\boldsymbol{\theta}}_{LS} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{N} (y(\boldsymbol{x}_i, \boldsymbol{\theta}) - z(\boldsymbol{x}_i))^2 \tag{4.1}$$

One common calibration method is ad hoc and adjusts the parameters by comparing selected metrics from the computer model and the corresponding observations. For example, the method used in [Chen and Zhuang, 2012] to calibrate the parameters is one of these methods. Another example is [Zhu and Zhuang, 2013], which applied a data assimilation method to calibrate the key parameters in Terrestrial Ecosystem Model (TEM). With the improvement of remote sensing technology, environmental data could be acquired with finer spatial and temporal resolutions. With the above approaches, it becomes computationally expensive, especially when adjusting the parameters dynamically with sequential data set is needed. To overcome this problem, people have developed efficient statistical methods to calibrate the param-

eters. In these methods, a statistical surrogate (*emulator*) is developed to substitute the computationally expensive computer models. Furthermore, people have developed a Bayesian framework to calibrate the unknown parameters (e.g. [Kennedy and O'Hagan, 2001, Conti et al., 2004]). This calibration approach is also called "inverse regression" and requires data from simulators as well as field observations. For computationally expensive computer models, an efficient *emulator* is needed in the process of Bayesian calibration of physical models.

In [Kennedy and O'Hagan, 2001], the authors thoroughly described the Bayesian calibration approach and also considered all sources of uncertainties to improve the discrepancy between the model and real process. [Williamson et al., 2013] applied and improved the Bayesian method by reducing the parameter space with history matching method. However, very few literatures work on sequential data, while in environmental science, the inputs and outputs are mostly time series data. In this situation, one approach is to have one model for each time point, but it will be very inefficient. [Conti and O'Hagan, 2010] suggests multi-output model to represent the time dependent outcome. For example, one year data of monthly NEP could be represented as 12 dimensional output. However, for a substantially long time period, the dimension of output will be very high. In Bayesian analysis, the curse-of-dimensionality has often been an issue. To overcome this issue, I will develop an efficient method to calibrate the dynamic computer model, specifically Terrestrial Ecosystem Model (TEM). Exploratory analysis shows that the outcome (e.g., GPP, NPP) presents strong seasonal pattern, which inspires us to develop statistical emulator which considers the the temporal pattern and dependency on input space separately.

4.2 Existing Framework: Bayesian Calibration

In this section, I will introduce the existing framework of Bayesian calibration method for physical models. The notations in this framework are described in Table

4.1. Suppose the unknown parameters to be calibrated are $\boldsymbol{\theta}$ and the known information (i.e, data) is $\boldsymbol{d}$, Bayesian inference of $\boldsymbol{\theta}$ is to compute conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{d}) \propto p(\boldsymbol{d}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta})$ is prior of $\boldsymbol{\theta}$ and $p(\boldsymbol{d}|\boldsymbol{\theta})$ is distribution of data conditional on $\boldsymbol{\theta}$. In the Bayesian framework, I will first work on the Gaussian process emulator (3.2) as I discussed in Chapter 3

$$\boldsymbol{y} = \eta(\boldsymbol{x}, \boldsymbol{\theta}) \sim N(\boldsymbol{m}_1(\boldsymbol{x}, \boldsymbol{\theta}), c_1((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta})))$$

where $m_1(\boldsymbol{x}, \boldsymbol{\theta}) = h_1(\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{\beta}_1$, $h_1(\boldsymbol{x}, \boldsymbol{\theta})$ is a vector of functions of $(\boldsymbol{x}, \boldsymbol{\theta})$ and $\boldsymbol{\beta}_1$ is the corresponding regression coefficients.

Table 4.1.: Variable notations

Notation	Description
$\boldsymbol{\theta}$	the true parameters to be estimated
$\boldsymbol{\vartheta}$	the design set of input parameters
$\boldsymbol{x}$	the input variables
$\boldsymbol{x}_i^*$	set of input variable for computer model
$\boldsymbol{x}_i$	set of observed input variables
$\boldsymbol{y}$	the output from computer model
$\boldsymbol{z}$	the observed output
N	number of design data set
n	number of observations
$\boldsymbol{t}_i$	set of calibration parameters

Next, I will work on the discrepancy between $\boldsymbol{z}(\boldsymbol{x})$ and $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\theta})$. In [Kennedy and O'Hagan, 2001], the discrepancy $\delta(\boldsymbol{x})$ is assumed to be independent of $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\theta})$. ϵ is the observational error. Then, the real output $\boldsymbol{z}$ consists of three parts: computer model emulator, discrepancy and measurement error

$$\boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\theta}) + \delta(\boldsymbol{x}) \tag{4.2}$$

$\delta(\boldsymbol{x})$ is a discrepancy model, for example, it is also assumed to be multivariate normal distribution

$$\delta \sim N(\boldsymbol{m}_2(\boldsymbol{x}), c_2(\boldsymbol{x}, \boldsymbol{x}'))$$

where $m_2(x) = h_2(x)^T\boldsymbol{\beta}_2$, $h_2(x)$ is a vector of functions of $\boldsymbol{x}$ and $\boldsymbol{\beta}_2$ is corresponding regression coefficients. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, $c_1(.,.)$ and $c_2(.,.)$ be covariance functions. Their corresponding hyperparameters in c_1 and c_2 are represented by ϕ_1 and ϕ_2. The prior information of $\boldsymbol{\theta}$ is independent of the others. The priors of hyperparameters are assumed to be

$$\begin{aligned} p(\boldsymbol{\beta}) &\propto 1 \\ p(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\phi}) &\propto p(\boldsymbol{\theta})p(\boldsymbol{\phi}) \end{aligned} \tag{4.3}$$

Let $D_1 = \{(\boldsymbol{x}_1^*, \boldsymbol{t}_1), ..., (\boldsymbol{x}_N^*, \boldsymbol{t}_N)\}$ and $D_2 = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ denote input data from computer model and field observations, H_1 and H_2 denote matrix $[h_1(\boldsymbol{x}_i^*, \boldsymbol{t}_i)]_{i=1:N}$ and $[h_2(\boldsymbol{x}_i^*)]_{i=1:n}$, respectively. Then, $E(\boldsymbol{y}) = H_1\boldsymbol{\beta}_1$ and $E(\delta) = H_2\boldsymbol{\beta}_2$. From equation (4.2), the expectation of $\boldsymbol{z}$ is

$$E(\boldsymbol{z}) = E(\boldsymbol{y} + \delta) = H_1\boldsymbol{\beta}_1 + H_2\boldsymbol{\beta}_2 \tag{4.4}$$

The variance matrix of $\boldsymbol{y}$ is V_1 which is $N \times N$ matrix, and the variance matrix of $\boldsymbol{z}$ is V_2 which is $n \times n$ matrix, C_1 is the covariance between $\boldsymbol{y}$ and $\boldsymbol{z}$. Then, the full set of data $\boldsymbol{d}^T = (\boldsymbol{y}^T, \boldsymbol{z}^T)$ is normal with the following mean and covariance

$$\begin{aligned} E(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\phi}) &= \boldsymbol{m}_d(\boldsymbol{\theta}) = H(\boldsymbol{\theta})\boldsymbol{\beta} \\ \boldsymbol{V}(\boldsymbol{d}|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\phi}) &= V_d(\boldsymbol{\theta}) \end{aligned}$$

Where

$$H(\boldsymbol{\theta}) = \begin{pmatrix} H_1 & 0 \\ H_1 & H_2 \end{pmatrix}$$

and

$$V_d(\boldsymbol{\theta}) = \begin{pmatrix} V_1 & C_1 \\ C_1^T & V_1 + V_2 \end{pmatrix}$$

With prior of $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\phi})$ in 4.3, the full posterior distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{d}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\phi})f(d; m_d(\boldsymbol{\theta}), V_d(\boldsymbol{\theta})) \tag{4.5}$$

$\boldsymbol{\beta}$ will be integrated out of (4.5), and hyperparameters $\boldsymbol{\phi}$ from $\boldsymbol{z}$ and $\boldsymbol{y}$ will be estimated. Then, the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d})$ is derived and be used to make inference about $\boldsymbol{\theta}$.

As discussed in Section 4.1, the multi-output emulation approach (e.g. [Conti and O'Hagan, 2010]) has dimensional issues when working not sequential outputs as in TEM. In the next section, I will introduce a new approach which will overcome this issue, and furthermore, I will provide an efficient method to estimate parameters and predict output with sequentially arrived data.

4.3 Proposed Approach

In [Kennedy and O'Hagan, 2001], a Bayesian emulation and calibration method has been discussed theoretically and applied on a toy model. This approach was also applied in their following literatures (e.g. [Conti et al., 2004]). Then, they extended the emulation method to multi-output emulation as in [Conti and O'Hagan, 2010]. In this literature, the application is on climate model which has time dependent output. The time varying output is assumed to be a multi-dimensional random vector which follows a multivariate normal distribution. The emulator is to surrogate the complex computer model, and is determined by the mean function $\boldsymbol{m}(\boldsymbol{x})$ and covariance function $c(\boldsymbol{x}, \boldsymbol{x}')$. In their application, the output has 12 time points, so the dimension of the emulator is 12. However, this multi-output emulation approach is limited in our application since the TEM model has time dependent outputs which arrived sequentially for many years. So the length of time is undetermined. For example, if there are 5 years data, the dimension of monthly output will be 60; if there are 6 years data, then the dimension will be 72. So the current approach becomes impractical for TEM. Secondly, [Conti and O'Hagan, 2010] did not extend the multi-emulation to calibration of model parameters. In another literature, [Liu and West, 2009] devel-

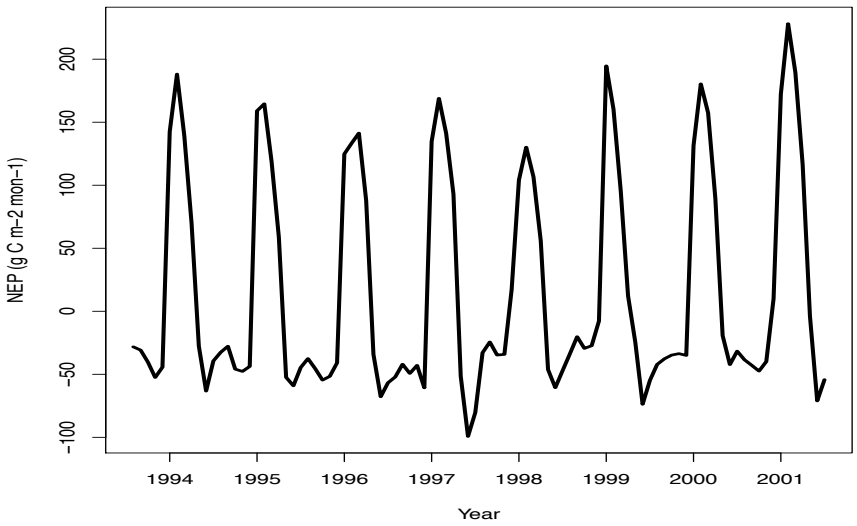

Figure 4.1.: Monthly NEP from 1994 to 2001

oped a dynamic emulator for computer models with time dependent output, which is also a Gaussian Process model and make time as a covariate. However, they did not extend the approach to calibration of input parameters. So in our new approach, I will develop a more efficient algorithm to surrogate the computer model and furthermore to calibrate parameters for the dynamic ecosystem model. In the next section, I will explore the real data and then develop a novel approach to study the TEM.

4.3.1 Motivation from Data

In TEM, the temporal scale is monthly time step for both input and output variables. The outputs of TEM include GPP, NEP and NPP. Here, I use NEP as an example. Figure 4.1 is the field observation of monthly NEP from 1994 to 2001 which has a clear seasonal pattern. In Figure 4.2, the trend magnitude of the data exhibits homogeneity monthly. Motivated by these facts, consideration of intra-annual variability would be a novel idea to build the emulator.

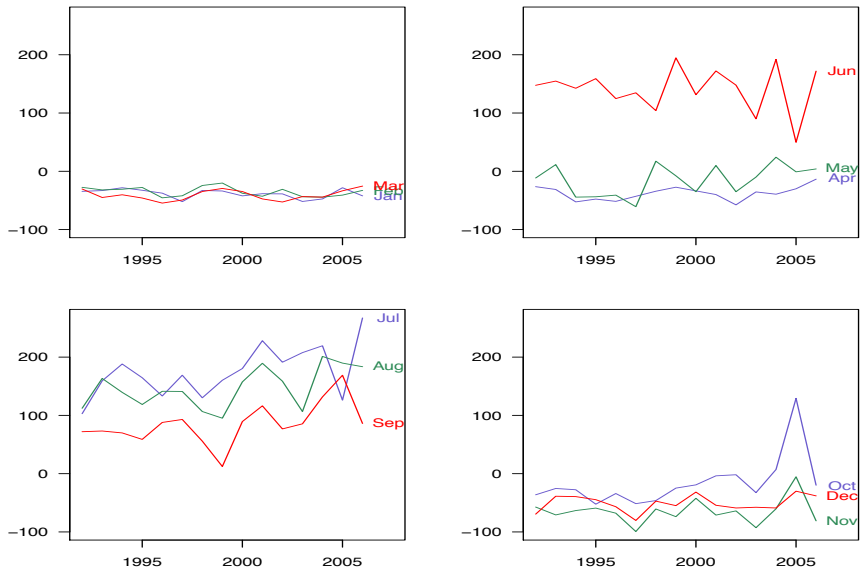

Figure 4.2.: BBplot of NEP from 1994-2001

The computer model output at time t is denoted by $y_t(\boldsymbol{x}, \boldsymbol{\theta})$. The field observations at t is $z_t(\boldsymbol{x})$. The correlation of outputs over time will become negligible when the lag becomes large. From this point of view, the dynamic linear model is an approximation to the high dimensional multi-output Gaussian process emulation.

In terms of the discrepancy between $y_t(\boldsymbol{x}, \boldsymbol{\theta})$ and $z_t(\boldsymbol{x})$, in [Kennedy and O'Hagan, 2001], it was modeled as $\delta(\boldsymbol{x})$ and independent of $\boldsymbol{\theta}$. Furthermore, [Kennedy and O'Hagan, 2001] assumes $\boldsymbol{\theta}(\boldsymbol{x})$ to be Gaussian process which has the same structure as $\boldsymbol{y}$. While in our approach, there will be two major differences: 1) the discrepancy model is assumed to be $\delta_t(\boldsymbol{x}, \boldsymbol{\theta})$ and depends on $\boldsymbol{\theta}$ because it is apparent that different value of $\boldsymbol{\theta}$ will result in different discrepancy data; 2) Model of $\delta_t(\boldsymbol{x}, \boldsymbol{\theta})$ is not necessary to be Gaussian process as in [Kennedy and O'Hagan, 2001] since there is zero uncertainty at training points. The task of modeling $\delta_t(\boldsymbol{x}, \boldsymbol{\theta})$ is a speed-accuracy tradeoff.

4.3.2 Computer Model Emulator: PAR with Gaussian Process

Our emulator is consisted of two parts: time series part which captures dependence and pattern across time and Gaussian process part which captures the dependence on inputs. Combination of the two will be the emulator which has zero uncertainty at training points and quantifies uncertainty at unknown points. It is a simpler surrogate of the computer model but computationally efficient for calibration of parameters and uncertainty quantification of dynamic models. Followed by the above discussion, the one dimensional complete model for z_t is

$$z_t(\boldsymbol{x}) = y_t(\boldsymbol{x}, \boldsymbol{\theta}) + \delta_t(\boldsymbol{x}, \boldsymbol{\theta}) \tag{4.6}$$

$y_t(\boldsymbol{x}, \boldsymbol{\theta})$ is time dependent computer output, and $z_t(\boldsymbol{x})$ is field observation at t. $\delta_t(\boldsymbol{x}, \boldsymbol{\theta})$ is the discrepancy.

Suppose the chosen parameter $\boldsymbol{\theta} = \boldsymbol{\vartheta}$ and $t = 1, \cdots, T$, the computer model output is

$$y_{1:T}(\boldsymbol{x}, \boldsymbol{\vartheta}) = (y_1(\boldsymbol{x}, \boldsymbol{\vartheta}), \ldots, y_T(\boldsymbol{x}, \boldsymbol{\vartheta}))$$

After seasonal pattern of NEP is explored, the periodic autoregression model (PAR) in [Pagano, 1978] is reviewed here.

$$Y_t = \sum_{j=1}^{p_t} a_{t,j} Y_{t-j} + \epsilon_t \tag{4.7}$$

with parameters $\phi_t = (a_{t,1}, ..., a_{t,p})$. When ϕ_t is constant over time, Y becomes AR(p) model. Lag $p_t = p$ be constant over time. $\epsilon_t \sim N(0, \sigma_t^2)$ is uncorrelated over time with mean zero. Assume period is s, then $E(\epsilon_t) = \sigma_t^2$, $\sigma_t^2 = \sigma_{t+s}^2$ and $\phi_t = \phi_{t+s}$.

Assuming input variables are observed at time t, $y_t(\boldsymbol{x}, \boldsymbol{\theta})$ is time series depending on $\boldsymbol{\theta}$. The univariate model emulator is

$$y_t(\boldsymbol{x}, \boldsymbol{\theta}) = \eta_t(\boldsymbol{\theta}) + M_1(\boldsymbol{x}, \boldsymbol{\theta}) \tag{4.8}$$

$$\eta_t(\boldsymbol{\theta}) = \sum_{j=1}^{p} a_{j,t} \eta_{t-j}(\boldsymbol{\theta}) + \epsilon_t \tag{4.9}$$

$$M_1(\boldsymbol{x}, \boldsymbol{\theta}) \sim GP(m_1(\boldsymbol{x}, \boldsymbol{\theta}), c_1((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta}))) \tag{4.10}$$

where $M_1(\boldsymbol{x}, \boldsymbol{\theta})$ is GP with mean $m_1(\boldsymbol{x}, \boldsymbol{\theta}) = h_1(\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_1$ is regression coefficients and $h_1(\boldsymbol{x}, \boldsymbol{\theta})$ is a vector of functions from set of variables $\{\boldsymbol{x}, \boldsymbol{\theta}\}$. For example, $h_1(\boldsymbol{x}, \boldsymbol{\theta}) = (1, \boldsymbol{x}, \boldsymbol{\theta})'$. Assuming dimension of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are d_1 and d_2 respectively, the covariance structure in $M_1(\boldsymbol{x}, \boldsymbol{\theta})$ is defined as

$$c_1((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta})) = \sigma^2 \exp\{-\sum_{l=1}^{d_1} r_l(x_{il} - x_{jl})^2 - \sum_{l=d_1+1}^{d_1+d_2} r_l(\theta_{il} - \theta_{jl})^2\} \tag{4.11}$$

where $\boldsymbol{r} = (r_1, \cdots, r_{d_1+d_2})'$. [Higdon et al., 2008] described the half-range method to estimate hyperparameter $\boldsymbol{r}$.

4.3.3 Estimation of Parameters in η_t

$\eta_t(\boldsymbol{\theta})$ is periodic autoregression of order p (PAR(p)) with coefficients $\phi_t = (a_{t,1}, ..., a_{t,p})$. To estimate ϕ_t which could be different over time, there have the following constraints when period is defined as s and $\{t = (n-1)s + m, n \in N^+, m = 1, \cdots, s\}$

1). $\phi_t = (a_{1,t}, ..., a_{p,t})$, $\phi_m = (a_{1,m}, ..., a_{p,m})$, $\phi_t = \phi_m$

2). $\epsilon_t(\boldsymbol{\theta})$ are uncorrelated with mean zero, variance σ_t^2

3). $E(\epsilon_t^2(\boldsymbol{\theta})) = \sigma_t^2$, $\sigma_t^2 = \sigma_{t+s}^2$ and $\phi_t = \phi_{t+s}$

In TEM, the period is one year, so $s = 12$ in the monthly temporal scale. $\eta_t(\boldsymbol{\theta})$ is nonstationary since the variance and covariance are different within a year. Next, I will discuss how to estimate the parameters ϕ_t. [Franses, 1996] discussed and implement how to estimate parameters in PAR(2) with period 4. In our study, I will introduce how to estimate PAR(3) with period 12.

To estimate the parameters ϕ_t, the first step is to determine the order of PAR. [Franses and Paap, 1994] discussed how to determine p using BIC in combination with diagnostic tests on residual autocorrelation. For example, if $p = 3$ and $\eta_t(\boldsymbol{\theta}) = \eta_{m,n}$, then in year n

$$
\begin{aligned}
m = 1: &\quad \eta_{1,n} = a_{1,1}\eta_{12,n-1} + a_{2,1}\eta_{11,n-1} + a_{3,1}\eta_{10,n-1} + \epsilon_{1,n} \\
m = 2: &\quad \eta_{2,n} = a_{1,2}\eta_{1,n} + a_{2,2}\eta_{12,n-1} + a_{3,2}\eta_{11,n-1} + \epsilon_{2,n} \\
m = 3: &\quad \eta_{3,n} = a_{1,3}\eta_{2,n} + a_{2,3}\eta_{1,n} + a_{3,3}\eta_{12,n-1} + \epsilon_{3,n} \\
m = 4: &\quad \eta_{4,n} = a_{1,4}\eta_{3,n} + a_{2,4}\eta_{2,n} + a_{3,4}\eta_{1,n} + \epsilon_{4,n} \\
m = 5: &\quad \eta_{5,n} = a_{1,5}\eta_{4,n} + a_{2,5}\eta_{3,n} + a_{3,5}\eta_{2,n} + \epsilon_{5,n} \\
m = 6: &\quad \eta_{6,n} = a_{1,6}\eta_{5,n} + a_{2,6}\eta_{4,n} + a_{3,6}\eta_{3,n} + \epsilon_{6,n} \\
m = 7: &\quad \eta_{7,n} = a_{1,7}\eta_{6,n} + a_{2,7}\eta_{5,n} + a_{3,7}\eta_{4,n} + \epsilon_{7,n} \\
m = 8: &\quad \eta_{8,n} = a_{1,8}\eta_{7,n} + a_{2,8}\eta_{6,n} + a_{3,8}\eta_{5,n} + \epsilon_{8,n} \\
m = 9: &\quad \eta_{9,n} = a_{1,9}\eta_{8,n} + a_{2,9}\eta_{7,n} + a_{3,9}\eta_{6,n} + \epsilon_{9,n} \\
m = 10: &\quad \eta_{10,n} = a_{1,10}\eta_{9,n} + a_{2,10}\eta_{8,n} + a_{3,10}\eta_{7,n} + \epsilon_{10,n} \\
m = 11: &\quad \eta_{11,n} = a_{1,11}\eta_{10,n} + a_{2,11}\eta_{9,n} + a_{3,11}\eta_{8,n} + \epsilon_{11,n} \\
m = 12: &\quad \eta_{12,n} = a_{1,12}\eta_{11,n} + a_{2,12}\eta_{10,n} + a_{3,12}\eta_{9,n} + \epsilon_{12,n}
\end{aligned}
$$

Let $\boldsymbol{E}_n = (\eta_{1,n}, \cdots, \eta_{12,n})^T$, where $\boldsymbol{E}_n$ is vector of quarters [VQ] representation. Then,

$$\Phi_0 \boldsymbol{E}_n = \Phi_1 \boldsymbol{E}_{n-1} + \boldsymbol{\epsilon}_n \tag{4.12}$$

with

$$\Phi_0(i,j) = \begin{cases} 1 & i = j \\ 0 & j > i \text{ or } i - j > 3 \\ -a_{i-j,i} & 1 \le i - j \le 3 \end{cases}$$

and

$$\Phi_1(i,j) = \begin{cases} 0 & j - i \le 8 \\ a_{i-j+12,i} & 9 \le j - i \le 11 \end{cases}$$

where $i, j = 1, 2, \cdots, 12$. The matrix form of Φ_0 and Φ_1 is in appendices.

For forecasting purpose, Φ_0^{-1} to (4.12) is applied and

$$\boldsymbol{E}_n = \Phi_0^{-1}\Phi_1\boldsymbol{E}_{n-1} + \Phi_0^{-1}\boldsymbol{\epsilon}_n \tag{4.13}$$

which is VAR(1). Φ_0^{-1} is a lower triangle matrix too.

To analysis the trend in η_t, the solutions is solved from the characteristic equation of (4.12)

$$|\Phi_0 - \Phi_1 z| = 0 \tag{4.14}$$

Suppose there are k unit root solutions of E_n process. E_n is stationary autoregression ([Pagano, 1978]).

To estimate the parameters $a_{m,n}$ and σ_m^2, there are N years data. Suppose the mean function of η_t is $m(t) = E(\eta_t)$, and the covariance kernel is

$$R(t_1, t_2) = E\{(\eta_{t_1} - m(t_1))(\eta_{t_2} - m(t_2))\} \tag{4.15}$$

For *periodically correlated* η_t, for all integers t_1 and t_2, then

$$m(t_1) = m(t_2), R(t_1, t_2) = R(t_1 + s, t_2 + s) \tag{4.16}$$

Without loss of generality, let $m(t) = 0$. Then, for integer $v \geq 0$,, multipling both sides of (4.9) by η_{t-v}, it becomes

$$\eta_t\eta_{t-v} \quad = \quad \sum_{j=1}^{p} a_{j,t}\eta_{t-j}\eta_{t-v} + \epsilon_t\eta_{t-v} \tag{4.17}$$

Take expectations of the obove equation, it becomes

$$R(t, t-v) = \sum_{j=1}^{p} a_{j,t}R(t-j, t-v) + \sigma_t^2 I_{v=0} \tag{4.18}$$

where I is identity function. Then, for $m = 1, 2, \cdots, 12$, $v \geq 0$,

$$R(m, m-v) = \sum_{j=1}^{p} a_{j,t}R(m-j, m-v) + \sigma_m^2 I_{v=0} \tag{4.19}$$

If there are N years data, $R(m, v)$ is approximated by

$$R_N(m, v) = \frac{1}{N}\sum_{j=0}^{k} \eta_{m+s*j}\eta_{v+s*j} \tag{4.20}$$

where $m = 1, 2, \cdots, s$, $v = 0, 1, \cdots, N * s - m - 1$, and $k = [N - \max(m, v)/s]$.

Replacing R by R_N in (4.19) and solving the linear equations, $a_{j,m}$ and σ_m^2 can be estimated.

4.3.4 Estimation of parameters in $M_1(\boldsymbol{x}, \boldsymbol{\theta})$

Assuming $d_2 = 1$ and $d_1 = 3$, there are n set of design points $\{\boldsymbol{\vartheta}_1, \cdots, \boldsymbol{\vartheta}_n\}$ and $\boldsymbol{x} = (x_1, x_2, x_3)'$ at $t = 1, \cdots, T$. The output $y_t(\boldsymbol{\vartheta}_j)$ are as follows:

$$\begin{aligned}
\boldsymbol{\vartheta}_1 : &\quad y_1(\boldsymbol{x}_1, \boldsymbol{\vartheta}_1), y_2(\boldsymbol{x}_2, \boldsymbol{\vartheta}_1), \cdots, y_T(\boldsymbol{x}_T, \boldsymbol{\vartheta}_1) \\
\boldsymbol{\vartheta}_2 : &\quad y_1(\boldsymbol{x}_1, \boldsymbol{\vartheta}_2), y_2(\boldsymbol{x}_2, \boldsymbol{\vartheta}_2), \cdots, y_T(\boldsymbol{x}_T, \boldsymbol{\vartheta}_2) \\
\cdots &\quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\boldsymbol{\vartheta}_n : &\quad y_1(\boldsymbol{x}_1, \boldsymbol{\vartheta}_n), y_2(\boldsymbol{x}_2, \boldsymbol{\vartheta}_n), \cdots, y_T(\boldsymbol{x}_T, \boldsymbol{\vartheta}_n)
\end{aligned}$$

After $a_{j,m}$ and σ_m^2 been estimated, $M_1(\boldsymbol{x}_t, \boldsymbol{\vartheta}_i) = \boldsymbol{y}_t(\boldsymbol{x}_t, \boldsymbol{\vartheta}_i) - \eta_t$ is calculated from (4.10), where $t = 1, \ldots, T$ and $j = 1, \ldots, n$. In principle, $h_1(\boldsymbol{x}, \theta)$ could be any function of $(\boldsymbol{x}, \theta)$. For example, if $h_1(\boldsymbol{x}, \theta) = (1, x_1, x_2, x_3, \theta)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$, then

$$H_1 = \begin{pmatrix} h_1(\boldsymbol{x}_1, \boldsymbol{\vartheta}_1) \\ \cdots \\ h_1(\boldsymbol{x}_T, \boldsymbol{\vartheta}_1) \\ h_1(\boldsymbol{x}_1, \boldsymbol{\vartheta}_2) \\ \cdots \\ h_1(\boldsymbol{x}_T, \boldsymbol{\vartheta}_2) \\ \cdots \\ h_1(\boldsymbol{x}_T, \boldsymbol{\vartheta}_n) \end{pmatrix} \quad \text{and} \quad \boldsymbol{M}_1 = \begin{pmatrix} M_1(\boldsymbol{x}_1, \boldsymbol{\vartheta}_1) \\ \cdots \\ M_1(\boldsymbol{x}_T, \boldsymbol{\vartheta}_1) \\ M_1(\boldsymbol{x}_1, \boldsymbol{\vartheta}_2) \\ \cdots \\ M_1(\boldsymbol{x}_T, \boldsymbol{\vartheta}_2) \\ \cdots \\ M_1(\boldsymbol{x}_T, \boldsymbol{\vartheta}_n) \end{pmatrix} \tag{4.21}$$

Now, $\boldsymbol{M}_1$ will be multivariate normal distribution as

$$\boldsymbol{M}_1 \sim MVN(H_1\beta, \Sigma) \tag{4.22}$$

Similar as in Section 3.2 and with reference to ([Oakley and O'Hagan, 2002]), let $\Sigma = \sigma^2 * A$ and $A = [\rho(M_1(i), M_1(j))]_{nT,nT}$. $COV(M_1(\boldsymbol{x}, \theta), M_1(\boldsymbol{x}', \theta')) = \sigma^2 c((\boldsymbol{x}, \theta) - (\boldsymbol{x}', \theta'))$ Conditional on $\boldsymbol{M}_1$, then

$$M_1(.)|\boldsymbol{\beta}, \Sigma, \boldsymbol{M}_1 \sim N(m_1^*(.), COV^*(.,.)) \tag{4.23}$$

where

$$\begin{aligned} m^*(\boldsymbol{x}, \theta) &= h_1(\boldsymbol{x}, \theta)^T \boldsymbol{\beta} + \boldsymbol{t}(\boldsymbol{x}, \theta)^T \Sigma^{-1} (\boldsymbol{M}_1 - H_1 \boldsymbol{\beta}) \\ COV^*(M_1(\boldsymbol{x}, \theta), M_1(\boldsymbol{x}', \theta')) &= COV(M_1(\boldsymbol{x}, \theta), M_1(\boldsymbol{x}', \theta')) - \boldsymbol{t}((\boldsymbol{x}, \theta)^T \Sigma^{-1} \boldsymbol{t}(\boldsymbol{x}', \theta') \\ \boldsymbol{t}((\boldsymbol{x}, \theta) &= (\rho(M_1(\boldsymbol{x}, \theta), M_1(\boldsymbol{x}_1, \boldsymbol{\vartheta}_1)), \cdots, \rho(M_1(\boldsymbol{x}, \theta), M_1(\boldsymbol{x}_T, \boldsymbol{\vartheta}_n))) \end{aligned}$$

The least squares estimation and variance of $\boldsymbol{\beta}$ is

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (H_1^T \Sigma^{-1} H_1)^{-1} H_1^T \Sigma^{-1} \boldsymbol{M}_1 \\ V(\hat{\boldsymbol{\beta}}) &= (H_1^T \Sigma^{-1} H_1)^{-1} \end{aligned}$$

Integrate our $\boldsymbol{\beta}$ from (4.23), then

$$M_1(.)|\Sigma \sim N(m_1^{**}(.), COV^{**}(.,.)) \tag{4.24}$$

where

$$\begin{aligned} m^{**}(\boldsymbol{x}, \theta) &= h_1(\boldsymbol{x}, \theta)^T \hat{\boldsymbol{\beta}} + \boldsymbol{t}(\boldsymbol{x}, \theta)^T \Sigma^{-1} (\boldsymbol{M}_1 - H_1 \hat{\boldsymbol{\beta}}) \\ COV^{**}(M_1(\boldsymbol{x}, \theta), M_1(\boldsymbol{x}', \theta')) &= COV^*(M_1(\boldsymbol{x}, \theta), M_1(\boldsymbol{x}', \theta')) + \\ &\quad (h_1(\boldsymbol{x}, \theta) - \boldsymbol{t}(\boldsymbol{x}, \theta)^T \Sigma^{-1} H_1)(H_1^T \Sigma^{-1} H_1)^{-1} * \\ &\quad (h_1(\boldsymbol{x}', \theta') - \boldsymbol{t}(\boldsymbol{x}', \theta')^T \Sigma^{-1} H_1)^T \end{aligned}$$

The *emulator* will be a combination of (4.13) and (4.24). Another issue is *periodic integration*. If (4.14) has unit root, then η_t is periodically integrated if there exist some a_m for $m = 1, \cdots, 12$ that $(1 - a_m B)\eta_t$ does not have unit root, where B is backward operator. I will address this issue in our application.

4.3.5 Computer Model Emulator: PR with Gaussian Process

Periodic regression (PR) is regression on sinusoidal functions. Suppose the period is s, for each $\boldsymbol{\vartheta}$, $\eta_t(\boldsymbol{\vartheta})$ defined in 4.8, it will be models as periodic regression as follows

$$\begin{aligned}\eta_t(\boldsymbol{\theta}) &= \alpha_0 + \alpha_1(\boldsymbol{\theta})sin(2\pi t/s) + \beta_1(\boldsymbol{\theta})cos(2\pi t/s) + \dots \\ &\quad \cdots + \alpha_p(\boldsymbol{\theta})sin(2\pi pt/s) + \beta_p(\boldsymbol{\theta})cos(2\pi pt/s) \end{aligned} \tag{4.25}$$

For each set of $\boldsymbol{\vartheta}_i$, the coefficients: $\alpha_j(\boldsymbol{\vartheta}_i)$ and $\beta_j(\boldsymbol{\vartheta}_i)$ are estimated. With all the N sets of $\alpha_j(\boldsymbol{\vartheta}_i)$ and $\beta_j(\boldsymbol{\vartheta}_i)$, $\alpha_j(\boldsymbol{\theta})$ and $\beta_j(\boldsymbol{\theta})$ will be interpolation from them. For example, $\hat{\alpha_j}(\boldsymbol{\theta}) = i_a + a_1 * \theta_1 + a_2 * \theta_2$ and $\hat{\beta}_j(\boldsymbol{\theta}) = i_b + b_1 * \theta_1 + b_2 * \theta_2$ which use linear regression models. This approach is an easier way to estimate the coefficients than PAR. The modeling of $M_1(\boldsymbol{x}, \boldsymbol{\theta})$ is the same as in previous section, but data M_1 will be calculated corresponding to the results from PR.

4.3.6 Computer Model Emulator: Uni-output Gaussian Process

Since each monthly NEP exhibits different trend and magnitude, the whole time period is divided to Jan, Feb, $\cdots$, Dec. Then, there will be twelve separate models as followed

$$y_j = \eta(\boldsymbol{x}, \boldsymbol{\theta})_j \sim N(m_1(\boldsymbol{x}, \boldsymbol{\theta}), c_j((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta}))), j = 1, \cdots, 12$$

where y_j is the emulator for the jth month, $m_1(\boldsymbol{x}, \boldsymbol{\theta})$ and $c_j((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta})$ are the corresponding mean and covariance functions. The estimation of coefficients and parameters will be the same as in [Kennedy and O'Hagan, 2001]. With all the twelve one-dimensional model, a complete emulator is a combination of them.

4.3.7 Computer Model Emulator: Multi-output Gaussian Process

A little revise to [Conti and O'Hagan, 2010], the output data is arranged by the 12 months and each year as a repeated measurement. In this way, the output $\boldsymbol{y}$ is

12 dimension, and each entry has multiple replicates (years). Then, the multi-output emulation will be

$$\boldsymbol{y} = \eta(\boldsymbol{x}, \boldsymbol{\theta}) \sim N(\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{\theta}), C((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta})))$$

where $\boldsymbol{y}$ is 12 dimension output. $\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{\theta})$ and $C((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta})$ are the corresponding mean and covariance function. The estimation of coefficients and parameters will be the same as in [Conti and O'Hagan, 2010].

4.4 Bias estimation

After the emulator developed for the dynamic computer model, I will develop a model to estimate the discrepancy between computer model and real process. In the framework of [Kennedy and O'Hagan, 2001], the authors assume $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\theta})$ is independent of bias $\delta(\boldsymbol{x})$. After examining the discrepancy between $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\theta})$ and $\boldsymbol{z}(\boldsymbol{x})$, δ depends on $\boldsymbol{\theta}$ obviously because different $\boldsymbol{\vartheta}_j, j = 1, \ldots, N$ results in different $\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\vartheta}_j)$, so does $\delta(\boldsymbol{x}, \boldsymbol{\vartheta}_j)$. So in our model, I assume the bias part is functions of $(\boldsymbol{x}, \boldsymbol{\theta})$. The input data for bias model is $(\boldsymbol{x}_t, \boldsymbol{\vartheta})$, and the corresponding bias data is given by

$$\delta_t(\boldsymbol{x}, \boldsymbol{\vartheta}_j) = \boldsymbol{z}_t(\boldsymbol{x}) - \boldsymbol{y}_t(\boldsymbol{x}, \boldsymbol{\vartheta}_i) \tag{4.26}$$

All the output up to time T is denoted by $\delta_{1:T}$. The structure of δ_t would be different depending on the discrepancy between $\boldsymbol{y}$ and $\boldsymbol{z}$.

(1). If the bias is not critical, the bias effect may be confounded with measurement error. Modeling of $\delta_t(\boldsymbol{x}, \boldsymbol{\theta})$ is not necessary.

(2). If the bias plays a crucial role, it needs to be estimated according to different patterns.

- If $\delta_t(\boldsymbol{x}, \boldsymbol{\vartheta})$ does not have temporal pattern, then

$$\delta_t(\boldsymbol{x}) \sim GP(m_2(\boldsymbol{x}), COV(\boldsymbol{x}, \boldsymbol{x}')) \tag{4.27}$$

– if $\delta_t(\boldsymbol{x}, \boldsymbol{\vartheta})$ exhibits temporal pattern

$$\begin{aligned} \delta_t(\boldsymbol{x}, \boldsymbol{\vartheta}) &= \zeta_t + M_2(\boldsymbol{x}, \boldsymbol{\vartheta}), t = p+1, ..., T & (4.28) \\ \zeta &= \sum_{j=1}^{p} b_{t,j} \zeta_{t-j}(\boldsymbol{x}) & (4.29) \\ M_2(\boldsymbol{x}, \boldsymbol{\vartheta}) &\sim GP(m_2(\boldsymbol{x}, \boldsymbol{\vartheta}), COV(\boldsymbol{x}, \boldsymbol{x}')) & (4.30) \end{aligned}$$

– if $\delta_t(\boldsymbol{x}, \boldsymbol{\vartheta})$ has seasonal pattern, the following predictive model is developed

$$\begin{aligned} \delta_t(\boldsymbol{\theta}) &= a_0 + a_1(\boldsymbol{\theta}) sin(2\pi t/s) + b_1(\boldsymbol{\theta}) cos(2\pi t/s) + \dots \\ &\quad \cdots + a_p(\boldsymbol{\theta}) sin(2\pi pt/s) + b_p(\boldsymbol{\theta}) cos(2\pi pt/s) & (4.31) \end{aligned}$$

where ζ is the temporal trend pattern of bias data. The estimation of coefficients in δ_t follow the same procedure as in previous sections as in (4.9) and (4.25).

4.5 Calibration and Prediciton

Calibration is to estimate the unknown parameters in underlying physical model by comparing field observations and computer model outputs. For example, in TEM, Table 5.3 gives the key parameters to be estimated in the computer model, and the observations from AmeriFlux tower. There are two sets of data set available: training data set in computer model $\mathcal{D}_1 = \{\boldsymbol{x}, \boldsymbol{\vartheta}, y_{1:T}\}$ and field observations $\mathcal{D}_2 = \{\boldsymbol{x}, z_{1:T}\}$.

The full set of data is $d_t^T = (\boldsymbol{y}_t, \boldsymbol{z}_t)^T$. After estimation of two sets of parameters in $\boldsymbol{y}_t$ and δ_t. The field observations z_t is given by (4.6). Then, the expectation of $\boldsymbol{z}$ is

$$\begin{aligned} E(z_t(\boldsymbol{x})|\mathcal{D}_1, \mathcal{D}_2) &= E(y_t(\boldsymbol{x}, \boldsymbol{\theta})|\mathcal{D}_1, \boldsymbol{\theta})) + E(\delta_t(\boldsymbol{x}, \boldsymbol{\theta})|\mathcal{D}_2) \\ &= \sum_{j=1}^{p} a_{t,j} \eta_{t-j} + M_1(\boldsymbol{x}, \boldsymbol{\theta}) + \delta_t(\boldsymbol{x}, \boldsymbol{\theta}) & (4.32) \end{aligned}$$

Hence the $N+1$ vector d_t

$$\begin{aligned} & E(d_t|\boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{\phi}_{1:s}, \psi_{1:s}, \Sigma, \nu_{1:s}) = E((y_t, z_t)^T|\boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{\phi}_{1:s}, \psi_{1:s}, \Sigma, \nu_{1:s}) \\ = & \begin{pmatrix} \sum_{j=1}^{p} a_{t,j} y_{t-j}(\boldsymbol{\theta}) + M_1(\boldsymbol{x}, \boldsymbol{\theta}) \\ \sum_{j=1}^{p} a_{t,j} y_{t-j}(\boldsymbol{\theta}) + M_1(\boldsymbol{x}, \boldsymbol{\theta}) + \delta_t(\boldsymbol{x}, \boldsymbol{\theta}) \end{pmatrix} & (4.33) \end{aligned}$$

Let V_1 be variance matrix of $\boldsymbol{y}_t$, and V_2 be variance of δ_t. Then, the variance of z_t will be $V_1 + V_2 + \sigma^2$. C_1 is the covariance between $\boldsymbol{y}_t$ and $\boldsymbol{z}_t$, which is estimated from the observations. Then, the variance matrix of d_t

$$V(d_t|\boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{\phi}_{1:s}, \psi_{1:s}, \Sigma, \nu_{1:s}) = \begin{pmatrix} V_1 & C_1^T \\ C_1 & V_1 + V_2 + \sigma^2 I_n \end{pmatrix} \tag{4.34}$$

With (4.33) and (4.35), the full distribution

$$p(d_t|\boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{\phi}_{1:s}, \psi_{1:s}, \Sigma, \nu_{1:s}) \sim MVN(E(d_t), V(d_t)|\boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{\phi}_{1:s}, \psi_{1:s}, \Sigma, \nu_{1:s}) \tag{4.35}$$

With previously estimated $\boldsymbol{r}, \boldsymbol{\phi}_{1:s}, \psi_{1:s}, \Sigma, \nu_{1:s}$ and prior of $\boldsymbol{\theta}$, the posterior of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|d_{1:T}) \propto p(\boldsymbol{\theta}) \coprod_{t=p+1}^{T} p(d_t; E(d_t|\boldsymbol{\theta}), V(d_t|\boldsymbol{\theta})) \tag{4.36}$$

The prediction of the outcome z_t at $T+1$ will be given by the prediction mean of $y_t(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$ and bias estimation δ_t from Section 5.4. The variance will be given by (4.35). Estimation of $\boldsymbol{\theta}$ is the inference from posterior distribution of $\boldsymbol{\theta}$.

5. APPLICATION TO AN ECOSYSTEM MODEL

5.1 Illustration of the Problem

In recent years, climate change has become a global environmental challenge facing the world. Greenhouse gases (e.g. CO_2, H_2O) play an important role in keeping the earth warm, but an increase of them might lead to warmer environment and climate change. After the industry revolution, human activities have led to increased emission of greenhouse gases. In the meantime, vegetation plants convert CO_2, H_2O and energy to food through photosynthesis. So the cycle of C and N is greatly influenced by human activities and vegetations. Many environmental scientists have already developed various versions of ecosystem models to simulate this cycle. Here, *Terrestrial Ecosystem Model* (e.g. [Zhu and Zhuang, 2014]) is presented as an example to illustrate the cycle:

$$\begin{aligned}
\frac{dC_V}{dt} &= GPP - R_A - L_C \\
\frac{dC_S}{dt} &= L_C - R_H \\
\frac{dN_V}{dt} &= NUPTAKE - L_N \\
\frac{dN_V}{dt} &= L_N - NETMIN \\
\frac{dN_{AV}}{dt} &= NINPUT + NETMIN - NLOST - NUPTAKE
\end{aligned}$$

where Table 5.1 describes the variables in above equations, and the equations have been solved with Runge-Kutta-Fehlberg method in [Fehlberg, 1969].

The TEM is driven by monthly data of temperature, precipitation and cloudiness as well as CO2 concentrations. Besides the vegetation variables, there are some fixed but unknown parameters, such as initial and boundary conditions, which influences the model outputs even with slightly change of the value. Uncertainty quantification

of these parameters and furthermore sensitivity analysis are carried out in many literatures (e.g. [Zhu and Zhuang, 2014, 2013], [Tang and Zhuang, 2009], [Chen et al., 2011]). From the sensitivity analysis in [Zhu and Zhuang, 2014], C_{MAX} and K_C are the two most sensitive parameters in TEM. In the following applications, C_{MAX} and K_C are assumed to be unknown and all the other parameters are fixed. Hence, $\boldsymbol{\theta} = (C_{MAX}, K_C)$, and $\boldsymbol{x}_t = (\text{temperature, precipitation, cloudiness})_t$ where t is monthly scale. The output interested in is $y_t = NEP_t = GPP_t - R_t$, where R_t is respiration by plants, heterotrophs and decomposers (the microbes).

Before the approaches from Chapter 4 applied to TEM, the inputs and outputs data in Section 5.2 will be explored. In Section 5.3, different types of emulator will be carried out and compared. Followed by Section 5.4, the bias model will be developed. In Section 5.5, the key parameters will be estimated and used to forecast future NEP. Furthermore, comparison of the results will be discussed in the last section.

5.2 Data

United States is covered by five major vegetation types: boreal, coniferous, deciduous, grassland and shrub land. For each vegetation type, the true $\boldsymbol{\theta}$ might be different. For example, C_{MAX} (maximum rate of photosynthesis) of grassland is different from that of deciduous. Generally, people assume that different vegetation types follow different models. In our application, the deciduous broadleaf forest model is chosen for application.

The field observation data is obtained from AmeriFlux level-4 data located at $(42.5°N, 72°W)$ which represents deciduous broadleaf forest land type. This data consists of half-hourly observed temperature, precipitation, cloudiness and NEP. They are aggregated to monthly scale in our application. The computer model data is from process-based TEM developed by *Ecosystems & Biogeochemical Dynamics Laboratory* at Purdue University. These two sets of data are both used in the modeling and validation process. From the TEM, $n_t = 30$ sets of training data and $n_v = 6$ sets

Table 5.1.: Variable in the differential equations of TEM

Notation	Type	Description
C_V	state variable	carbon in vegetation
C_S	state variable	carbon in soil
N_V	state variable	nitrogen in vegetation
N_S	state variable	nitrogen in soil
N_{AV}	state variable	nitrogen in detritus
GPP	fluxes	gross primary production
NEP	fluxes	net ecosystem production
R_A	fluxes	autotrophic respiration
L_C	fluxes	carbon in litters
R_H	fluxes	heterotrophic respiration
L_N	fluxes	nitrogen in litters
$NINPUT$	fluxes	nitrogen input from outside ecosystem
$NETMIN$	fluxes	net rate of mineralization of N_S
$NLOST$	fluxes	nitrogen losses from ecosystem
$NUPTAKE$	fluxes	nitrogen uptake by vegetation

of validation data are simulated. For each set, monthly NEP from 1992 to 2006 are simulated from TEM. The input variables are field observations of monthly temperature, precipitation and cloudiness from 1992 to 2006. In the field observation, there are missing data (January and February in 2002), and bootstrap and multiple linear regression model are used to imputed the missing data.

Table 5.2.: Input variables in Terrestrial Ecosystem Model

$\boldsymbol{x}$	Description	temporal scale	unit
TEMP	surface air temperature	monthly	^oC
PREC	precipitation	monthly	mm
CLDINESS	cloudiness	monthly	%

5.2.1 Model Inputs and Parameters

The input variables $\boldsymbol{x}$ are described in Table 5.2. They are time series data and obtained from field observations. Figure 5.1 shows the field observations of monthly temperature, precipitation and cloudiness between 1992 and 2006 at $(42.5°N, 72°W)$. It is obvious that monthly temperature exhibits seasonal pattern, while the patterns of precipitation and cloudiness are less obvious.

Next, I will introduce the parameters $\boldsymbol{\theta}$. As discussed in previous sections, C_{MAX} and K_C are the two most sensitive parameters ([Zhu and Zhuang, 2014] and described in Table 5.3. The range of them refers to [Zhu and Zhuang, 2013].

Table 5.3.: Most sensitive parameters for deciduous broadleaf forest

ID	Acronym	Definition [a]	Prior [b]	Units
θ_1	C_{MAX}	Maximum rate of photosynthesis C	$[1200, 1700]$	$gm^{-2}mon^{-1}$
θ_2	K_C	Half saturation constant for CO_2 uptake by plants	$[200, 500]$	μLL^{-1}

[a] Reference: [Zhu and Zhuang, 2014].

[b] Referred to [Zhu and Zhuang, 2014] and [Zhu and Zhuang, 2013].

After input variables and parameters are setup, the TEM is run for 36 times and the monthly NEP is simulated for both training and validating sets. The prior of parameters are based on previous research results (e.g. [Zhu and Zhuang, 2013]).

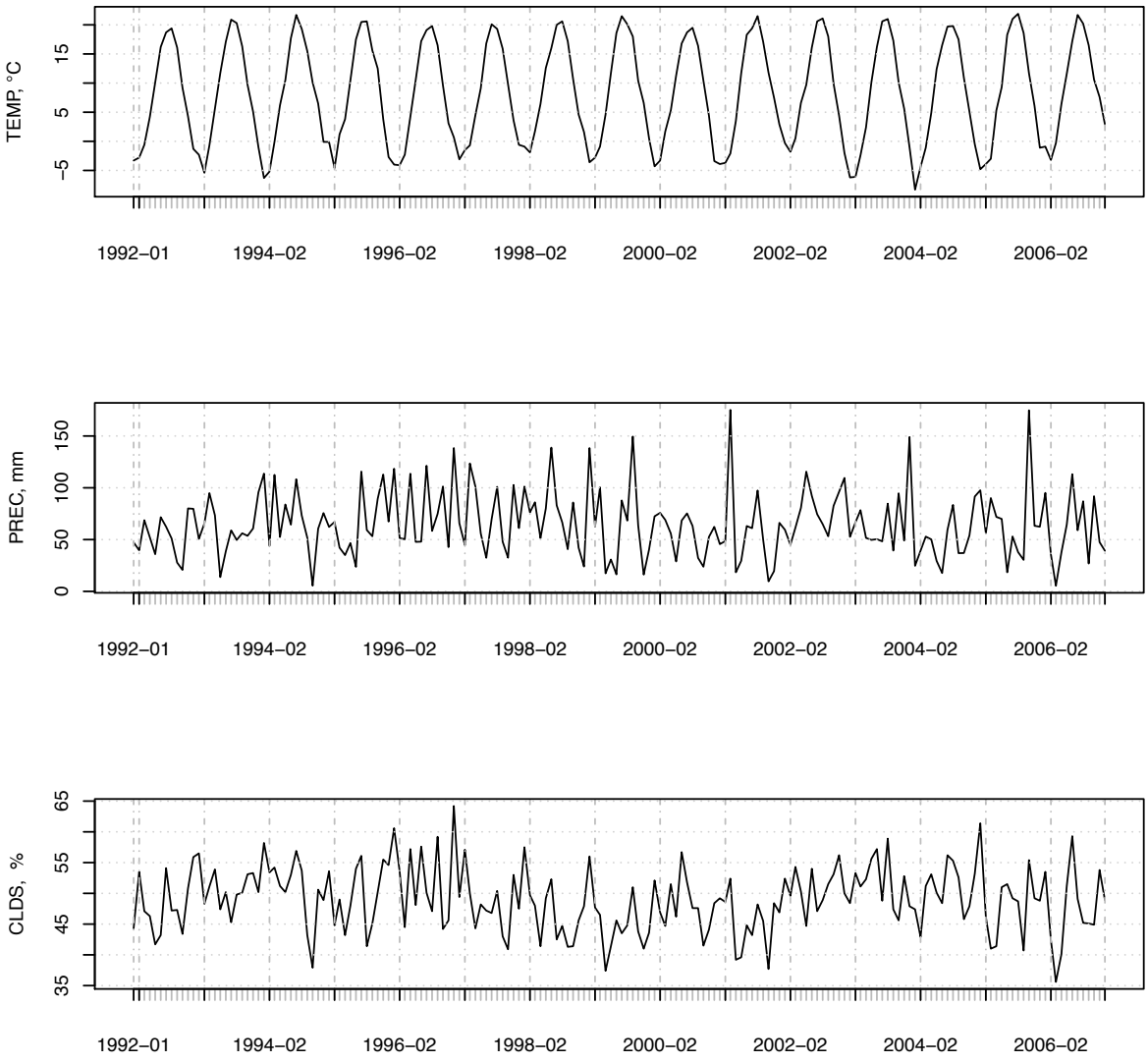

Figure 5.1.: Observations of monthly TEMP, PREC, and CLDS

5.2.2 Model Outputs and Field Observations

The output NEP is the difference between GPP and respiration by plants. Usually, people estimate GPP and respiration separately, and NEP is derived from them. To understand NEP better, the field observations are explored. Figure 5.2 shows the

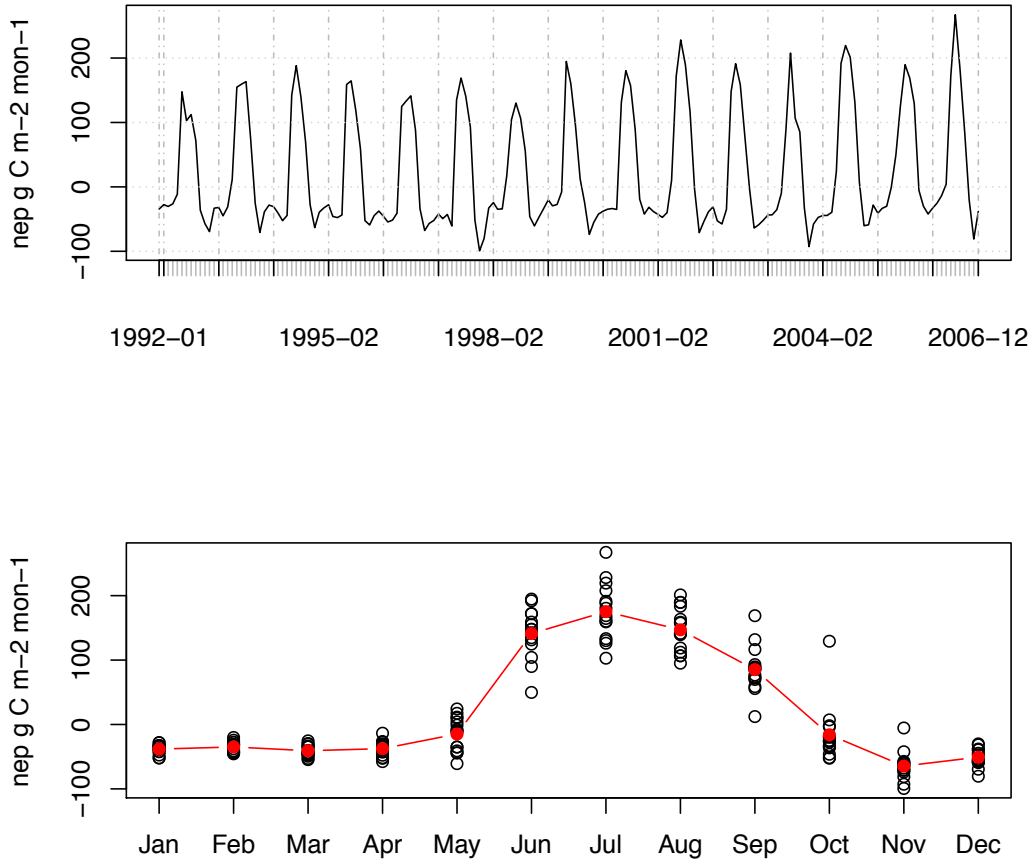

Figure 5.2.: Field observations of NEP between 1992 and 2006

monthly NEP from 1992 to 2006. From this graph, the seasonal pattern of NEP is very obvious and NEP reaches high peak during summer. Also, within month variations from May to September are relatively higher than that of other months, which implies a model considering the difference among months. Figure 5.3 further examines the variation of monthly NEP and presents the difference. I will consider two ways to take into account the within month variation of NEP: 1) different model for each month; 2) removal of seasonal pattern.

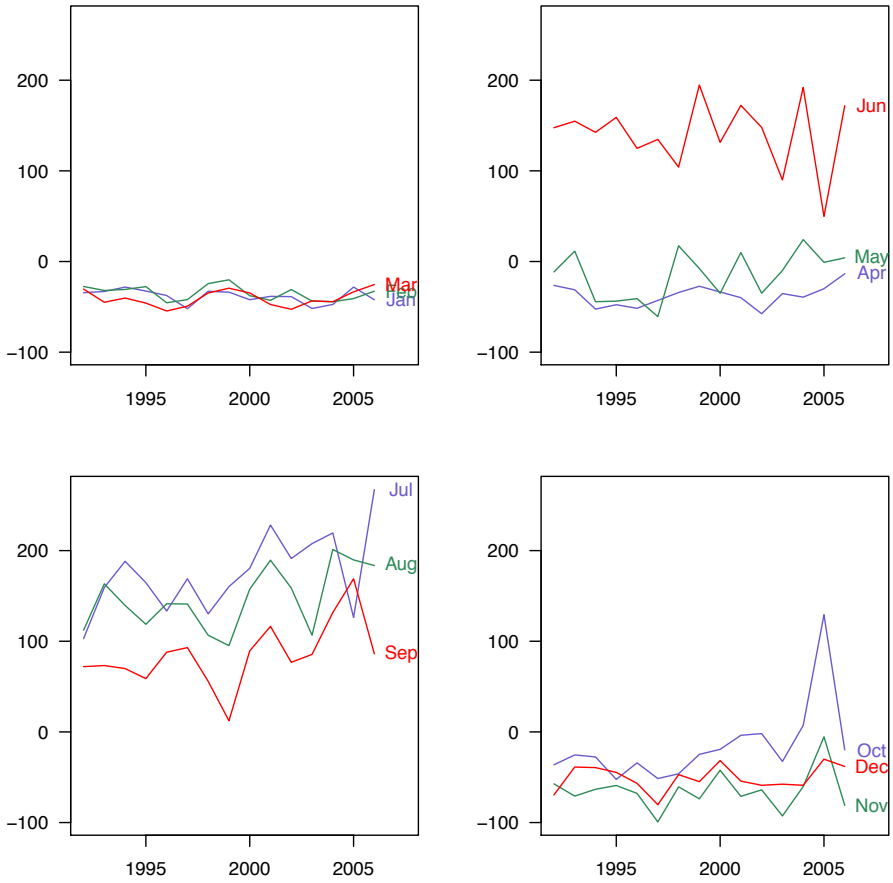

Figure 5.3.: Field observations of monthly NEP ($gCm^{-2}mon^{-1}$)

Next, monthly NEP from TEM which depends on input variables $\boldsymbol{x}$ and chosen parameter $\boldsymbol{\theta}$ values are explored as well. With uniform prior of $\boldsymbol{\theta}$: $[1200, 1700] \times [200, 500]$, 36 sets of parameter values are designed using LHS method, of which 6 sets are for validation purpose. Figure 5.4 shows the training points (black) and validation points (red) which span the whole input space.

In Figure 5.5, a parameter value $(1606, 272)$ is chosen, and the monthly NEP from model (black) with field observations(blue) are compared. The bias (red) between them is on the right and not negligible as shown in Figure 5.5. This further suggests that bias is not just noise, and there should be a model to estimate it. In the following

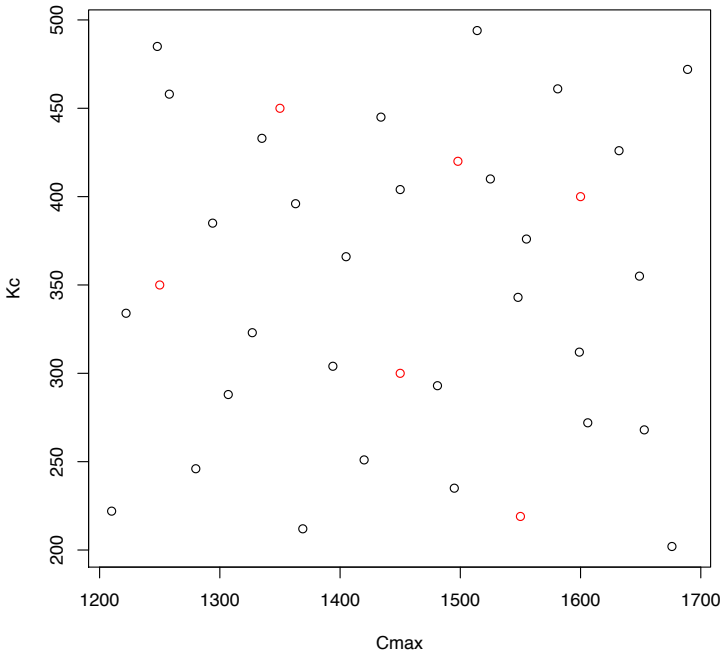

Figure 5.4.: Training set (black) and validation set (red) of $\boldsymbol{\theta}$

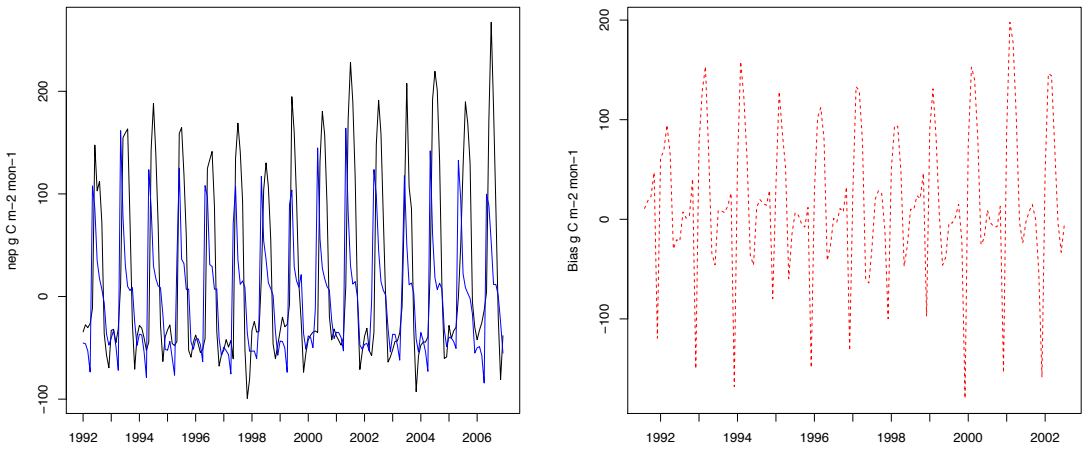

Figure 5.5.: NEP: field observations (black) vs. simulator (blue) vs bias (red)

sections, the emulator for TEM will be developed and furthermore to estimate $\boldsymbol{\theta}$. After that, I will proceed to forecast future NEP from the calibrated models.

5.3 Computer Model Emulators

In this section, emulators for TEM are developed from different methods as described in Chapter 4. Afterwards, I will compare the model accuracy of these three methods: 1) One dimensional emulation as in [Kennedy and O'Hagan, 2001]; 2) Multi-Output emulation as in [Conti and O'Hagan, 2010]; 3) PR with Gaussian process as in Chapter 4. Table 5.4 gives the notation and brief comparison of the three different methods.

Table 5.4.: Comparison of three emulation methods

Methods	Models		Dimension			
	Notation	Description	$\boldsymbol{\theta}$	$\boldsymbol{x}$	t	$\boldsymbol{y}$
1	Uni_GP	Uni-output emulator	2	3	0	1
2	Multi_GP	Multi-output emulator	2	36	0	12
3	PAR_GP	Time dependent emulator	2	3	1	1

After the emulators built, I will evaluate how well they represent the computer model. The three methods to be evaluated and compared are described in Table 5.4, and the measurements for comparison are $RMSE$, error rate and R-square defined as follows.

- Data
  - Training samples of $\boldsymbol{\theta}$: $\mathcal{S}_1 = \{\boldsymbol{\vartheta}_1, \cdots, \boldsymbol{\vartheta}_{30}\}$. For each sample point, monthly NEP ranging from 1992-2002 from TEM are simulated.
  - Validation samples of $\boldsymbol{\theta}$: $\mathcal{S}_2 = \{\boldsymbol{\vartheta}_{31}, \cdots, \boldsymbol{\vartheta}_{36}\}$. For each sample point, monthly NEP ranging from 1992-2002 are simulated from TEM .

- Emulation
  - Uni-output: for each month (Jan - Dec), $j = 1, 2, \cdots, 12$, there is

$$y_j \sim GP(m_j(\boldsymbol{x}, \boldsymbol{\theta}), c_j((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta}')) \tag{5.1}$$

    The emulation is combined by 12 models. y_j is j^{th} monthly NEP and depend on $\boldsymbol{\theta}$ and $\boldsymbol{x}$.

  - Multi-output: for one years monthly NEP, it becomes

$$\boldsymbol{y} \sim GP(\boldsymbol{m}(\boldsymbol{x}, \boldsymbol{\theta}), C((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta}')) \tag{5.2}$$

    where $\boldsymbol{y}$ is 12-dimensional (or 24-dimensional for two years). Each dimension of $\boldsymbol{y}$ is one month of NEP. The covariance function $C((\boldsymbol{x}, \boldsymbol{\theta}), (\boldsymbol{x}', \boldsymbol{\theta}') = \sigma^2\Sigma$.

  - PAR with GP: it will follow the equations (4.8) and (4.9), which is a trend model and a GP model.

- Validation
  - $RMSE_j = \sqrt{\sum_{i=1}^{n_v}(\hat{y_{ij}} - y_{ij})^2/n_v}$
  - $error\% = RMSE_j/SE_j$
  - R^2 = R-square of linear regression model: $y_j \sim a_j\hat{y}_j + \epsilon_i$, a_j is scaler.

5.3.1 Compare Different Emulators

In this section, I will compare the new approach with existing approaches in terms of accuracy and efficiency. In the new approach, we present the results from PR_GP which results in more accurate prediction than PAR_GP.

Table 5.5 is the evaluation results for each monthly NEP. In this table, the RMSE of PR_GP is slightly higher than that of Uni_GP, but both are smaller than that of Multi_GP. Furthermore, I compared the annual NEP from three different emulation

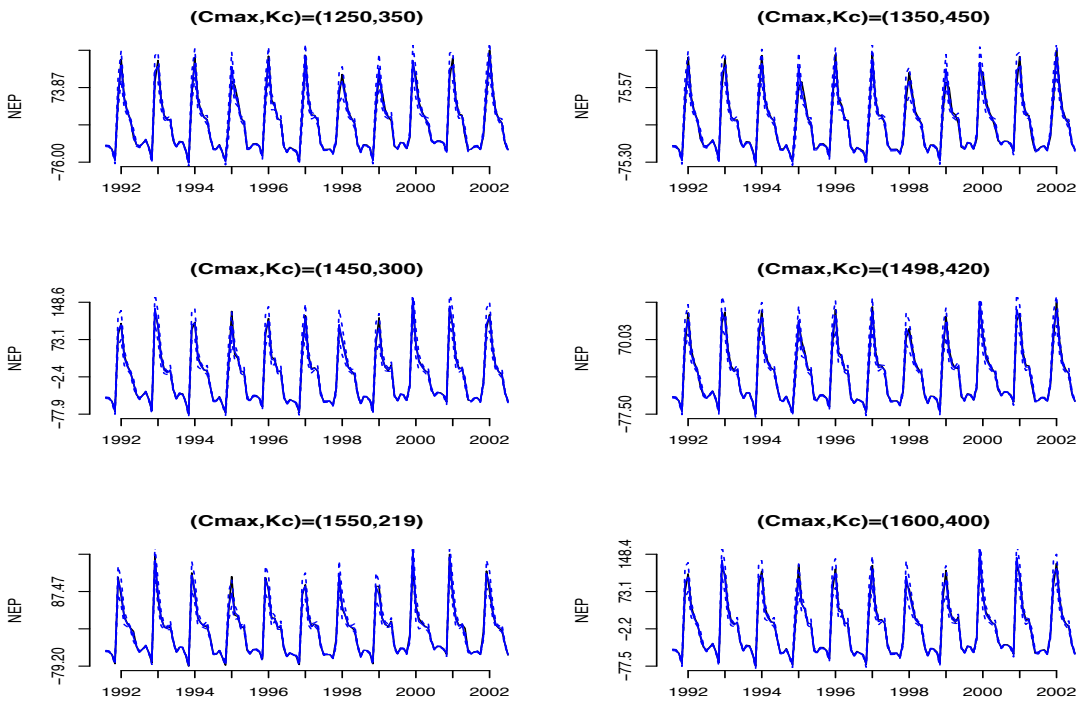

Figure 5.6.: PR_GP emulator on validation set

methods as in Table 5.6. The results reinforces that Uni_output model has the best emulation accuracy, PR_GP follows and Multi_GP is worst. However, since Uni_GP has 12 models, the computation time of Uni_GP is much longer than PR_GP. Table 5.6 further compares the annual RMSE from each model. Overall, Uni_GP gives the best model accuracy but take the most time. PR_GP works slightly worse than Uni_GP in terms of model accuracy, but is more computational efficient.

5.4 Bias Models

In this section, I will build the model to predict bias between the computer model and field observations. The field observations does not depend on $\boldsymbol{\theta}$, while the computer model depends on $\boldsymbol{\theta}$. Since bias is the difference between computer model and field observations, it also depends on $\boldsymbol{\theta}$. This is different from [Kennedy and

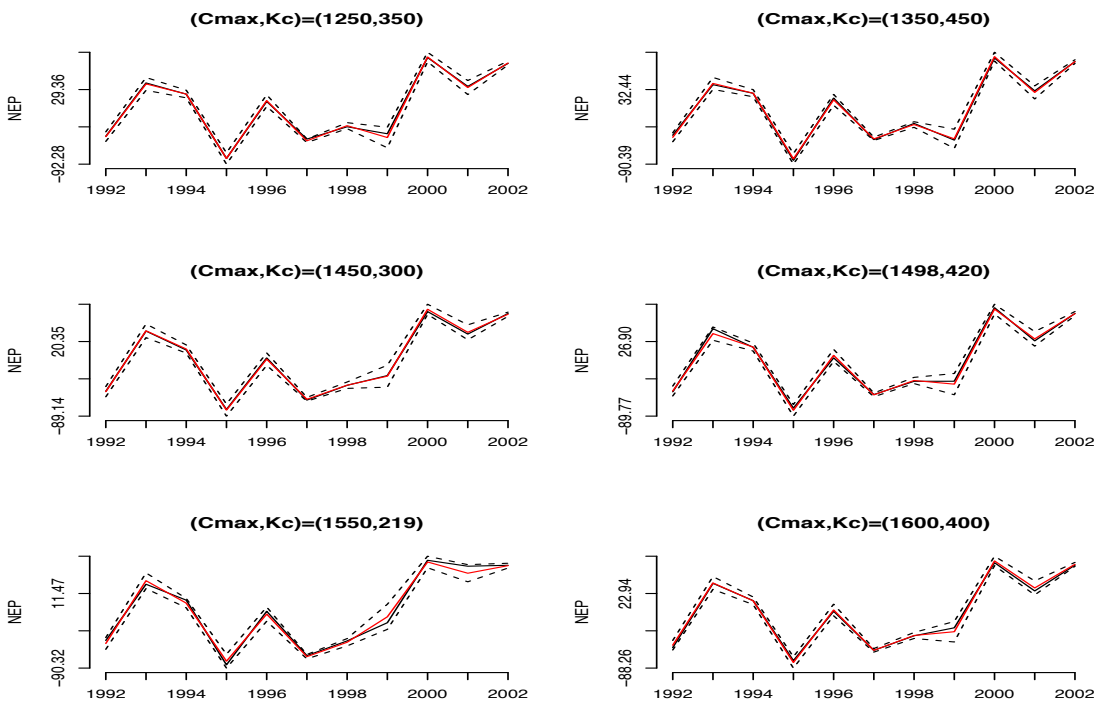

Figure 5.7.: Comparison of Emulation on yearly scale NEP

O'Hagan, 2001] which assumes bias is independent of $\boldsymbol{\theta}$. The bias model is not an emulator since there is no complex computer model for bias, and no need for zero uncertainty at training points. From the bias plot as shown in Figure 5.5 and 5.8, it is larger in growing season than non growing season. Multiple regression model on $(\boldsymbol{x}, \boldsymbol{\theta})$ with seasonal trend best fits the bias data. With the fitted model, the observed and predicted bias are shown in Figure 5.9 from one parameter value.

5.5 Calibration and Forecasting

The predictions of future NEP will be bias corrected with $\delta_t(\boldsymbol{x})$ and compared with the real observations z_t.

Table 5.5.: Comparison of RMSE with Emulation for each month

Month	Uni_GP	Multi_GP	PR_GP
1	0.783	1.880	0.781
2	0.776	1.880	0.782
3	0.843	2.160	0.863
4	2.010	2.840	3.475
5	15.290	13.430	15.979
6	12.500	16.840	14.903
7	7.680	7.870	8.253
8	3.220	3.420	3.408
9	1.630	2.210	1.797
10	3.630	4.360	4.762
11	1.570	1.630	1.855
12	0.840	2.390	0.851

Table 5.6.: RMSE of annual NEP comparison

Emulations	Simulation		Monthly		Annual	
	n_t	n_v	RMSE(M)	$error.rate(\%)$	RMSE(M)	$error.rate(\%)$
Uni_GP	30	6	6.35	11.30	5.86	11.40
Multi_GP	30	6	8.65	18.60	15.80	30.90
PR_GP	30	6	7.08	12.60	7.49	16.5

After the emulator of computer model and bias model developed, the distribution of $p(\boldsymbol{\theta}|D_1, D_2)$ is derived. Then, Metropolis algorithm is used to estimate the parameters. Here is the algorithm.

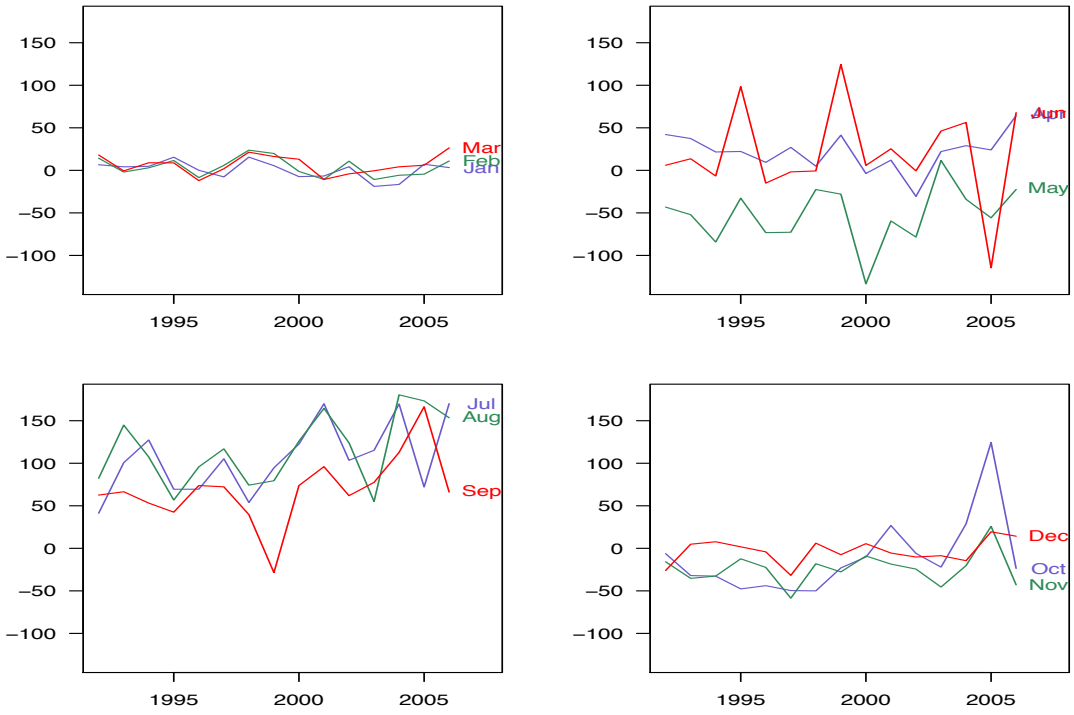

Figure 5.8.: Monthly Bias of NEP from 1992 to 2006

Step 1: Likelihood of function of data

$$\begin{aligned} p(d_i|\boldsymbol{\theta}) &\propto \frac{1}{\sqrt{V(d_i)}}\dot{\exp}\{-\frac{(d(\boldsymbol{\theta})-d_i)^2}{2V(d_i)}\} \\ log(p(d_i|\boldsymbol{\theta})) &= -0.5\frac{(d(\boldsymbol{\theta})-d_i)^2}{V(d_i)} - log(\sqrt{V(d_i)}) + const \\ log(Likelihood|\boldsymbol{\theta}) &= \Sigma_{i=1}^{T} log(p(d_i|\boldsymbol{\theta})) + const \end{aligned} \tag{5.3}$$

Step 2: Posterior distribution of $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\boldsymbol{d}) \propto p(\boldsymbol{\theta})\Pi_{i=1}^{T} p(d_i|\boldsymbol{\theta})$$

Step 3: Sample from posterior distribution of $\boldsymbol{\theta}$ with Metropolis algorithm

1). Select the starting point in the parameter space, eg.

2). Calculate the likelihood of the prior

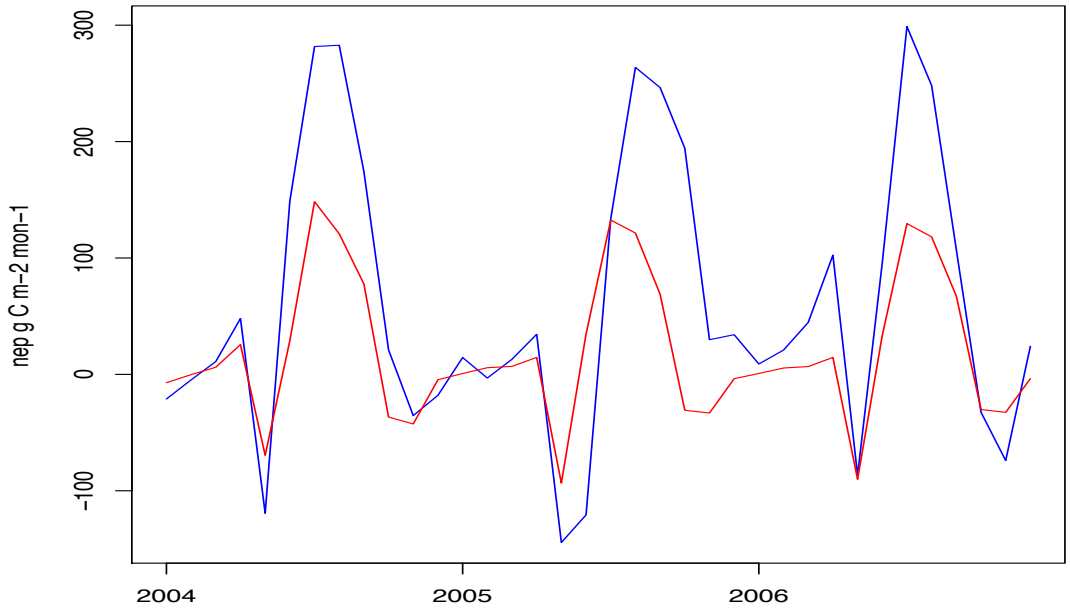

Figure 5.9.: Observed vs. Predicted Bias (blue) between 2004 and 2006

3). Set i=1

4). Run the following: (i) Generate the new point (ii) Calculate the new likelihood (iii) Calculate the Metropolis ratio (iv) Accept the candidate with probability equal to min(Ratio, 1) (v) If the candidate is accepted (vi) Set i= i +1

5). A representative sample from the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d})$.

Before making predictions, the parameters will be estimated. The posterior distribution of the parameters are shown in Figure 5.10. The 2 dimensional plot of posterior samples is shown in Figure 5.11. From the posterior distribution, the $\boldsymbol{\theta}$ is estimated.

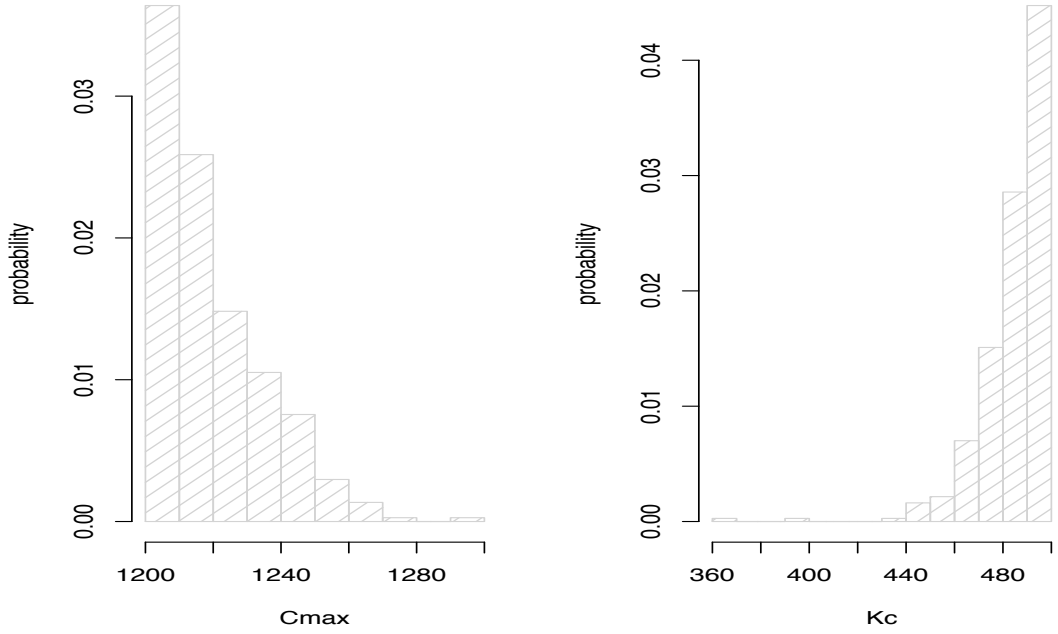

Figure 5.10.: Posterior distribution of C_{MAX} and K_C

Figure 5.12 gives us the histogram of NEP and fitted NEP from our final model. The residual looks normal, so it would be reasonable to assume Gaussian process on the residuals.

Table 5.7.: Comparing predictive accuracy in 2003-2006

Methods	Estimation		Forecasting		
	$\hat{C}_{MAX}$	$\hat{K}_C$	MSE	SCALE	R^2
Uni-output	1231.00	495.00	52.3	0.972	0.725
Multi-output	1231.00	495.00	83.3	0.606	0.303
PR with GP	1231.00	495.00	**41.3**	**1.06**	**0.828**
[Zhu and Zhuang, 2014]	1498.00	420.00	73.2	1.21	0.461
[Zhu and Zhuang, 2013]	1141.02	219.83	74.3	1.20	0.444

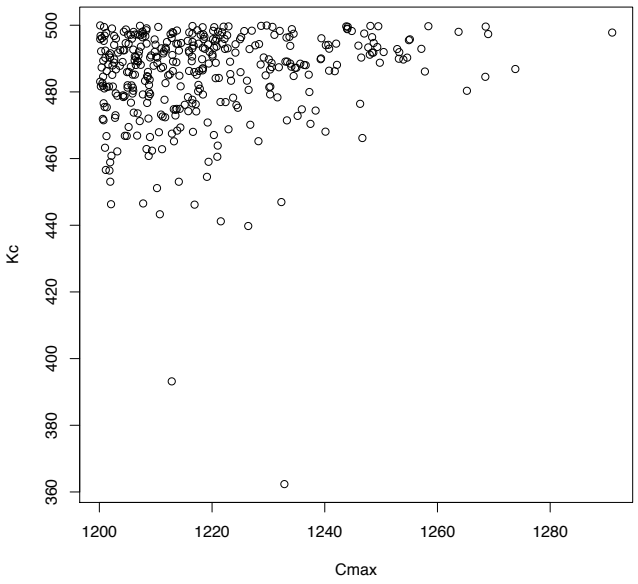

Figure 5.11.: Samples from posterior distribution of (C_{MAX}, K_C)

Table 5.7 shows the comparison of precision with different methods. R^2 is the R-square of field observed NEP regressed on predicted NEP. Figure 5.13 shows the comparison of forecasting in 2004-2006 from three different methods. The PR_GP gives the best prediction results in terms of R-square and MSE. From the computer model dat and field observations, I constructed a new relationship between $\boldsymbol{x}$, $\boldsymbol{\theta}$ and $\boldsymbol{y}$ as in Figure 5.13. The predicted NEP vs field observation is adding up Figure 5.9 and Figure 5.15.

5.6 Conclusion and Discussion

In this chapter, I have developed a new dynamic emulator as a surrogate for TEM. This approach consists of PR and GP, where the PR part captures the time varying effect of inputs, and the GP captures the correlation structure among input variables. To determine which model structure to choose for the emulator and bias correction

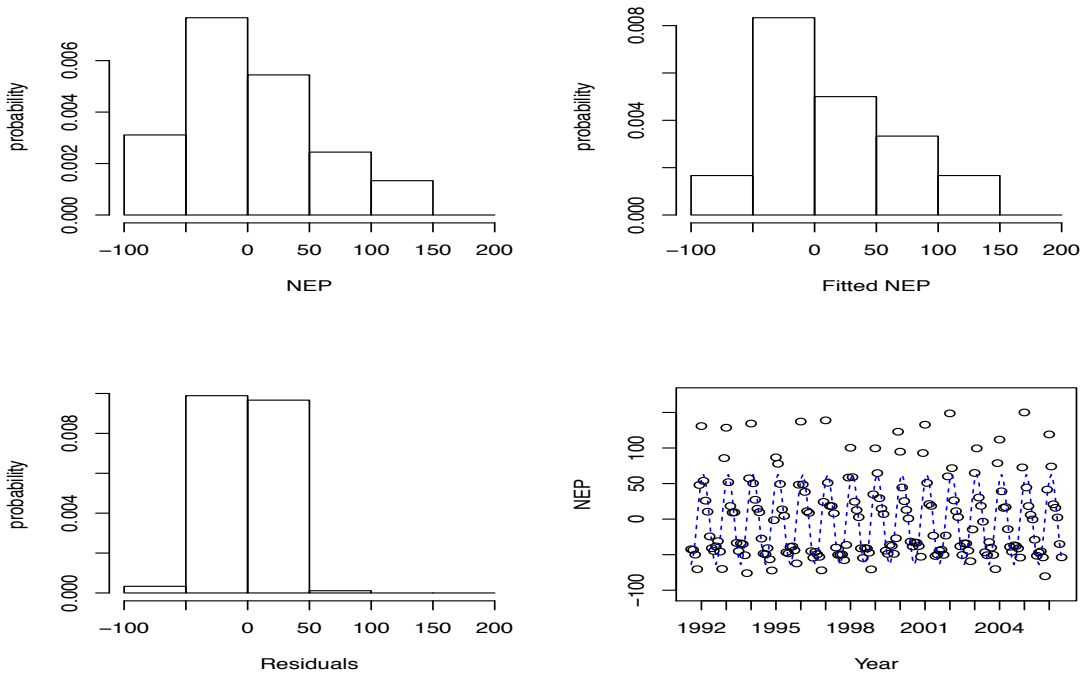

Figure 5.12.: Histograms

part, I explored the real observations carefully at first. Next, I compared the real observations with computer model output and assessed the difference.

With small number of training data set, our emulator works well with the dynamic computer model. The bias model was built based on the difference between computer model output and field observations. Different from the existing approach, I have considered to model the bias while varying the parameters, and this improvement helps us to make better predictions. Furthermore, I extended the emulation to calibrate the computer model and make predictions of future NEP after I built bias model. Finally, I evaluated our method with comparison to existing approaches with a set of validation data. Compared with the existing approach interns of forecasting future NEP, our approach works the best in prediction accuracy and efficiency.

Overall, Bayesian calibration method is an effective approach to calibrate ecosystem models. With available data, I could estimate the unknown parameters along

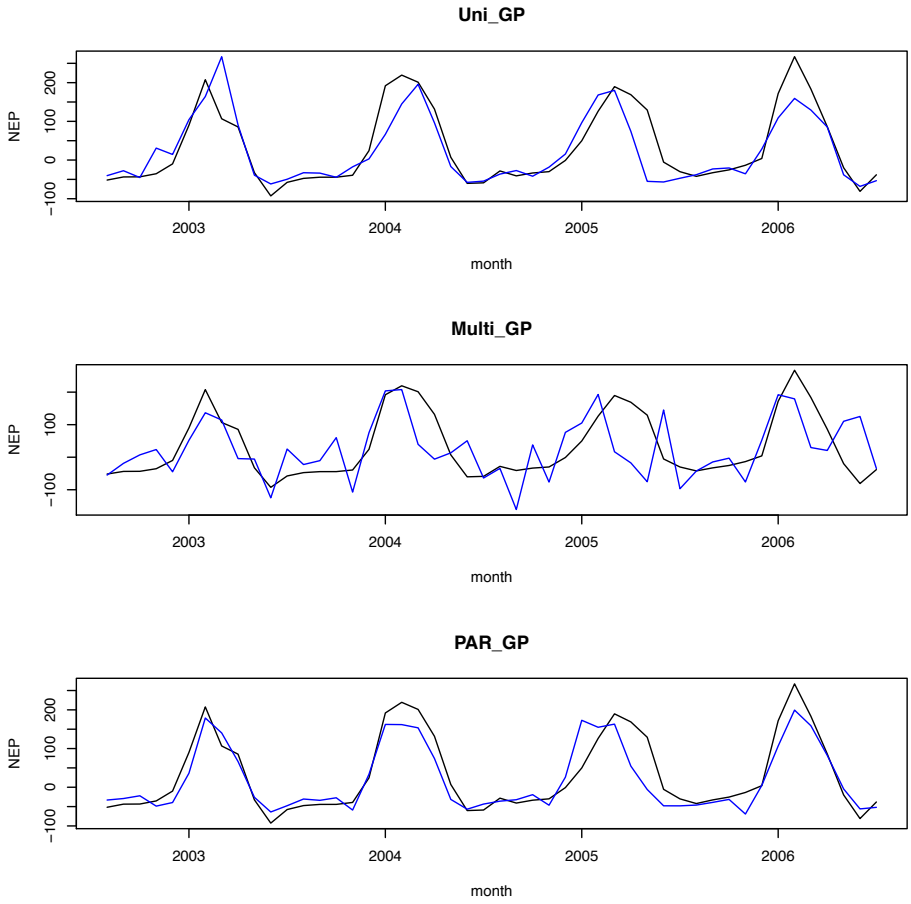

Figure 5.13.: Field observation (black) vs Predicted NEP (blue)

with their uncertainties. However, for a specific computer model, I need to investigate the model itself and explore the real data thoroughly to make assumptions on the emulator structure. In the future, I will extend current work to regional level with all the five vegetation types: deciduous, boreal, coniferous, grassland, shrub land. Then, I could evaluate our results in regional level. Furthermore, I will assess other important carbon fluxes (e.g., gross primary production) as an additional output.

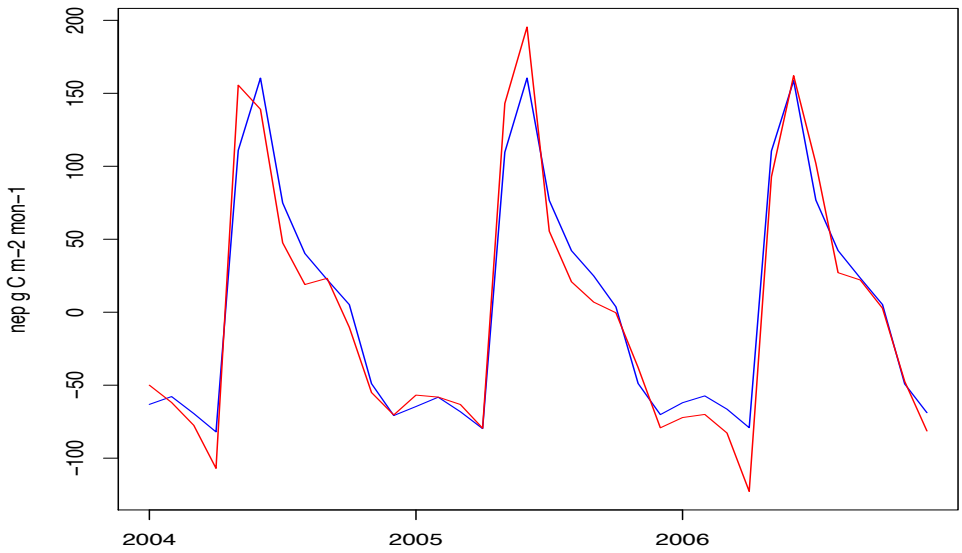

Figure 5.14.: Model output (blue) vs. Emulator (red) between 2004 and 2006

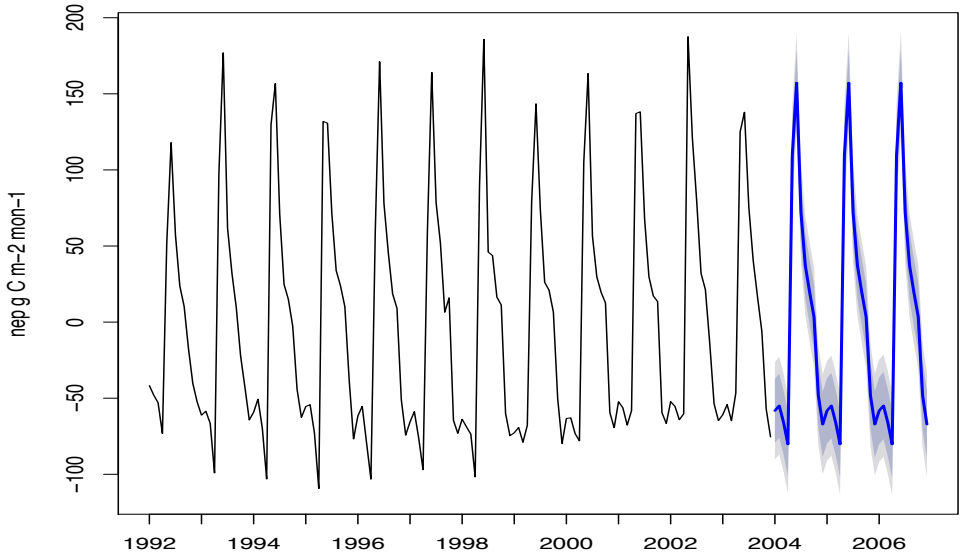

Figure 5.15.: Predicted model output (blue) between 2004 and 2006

APPENDICES

Appendix A

$$\Phi_0 = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-a_{1,2} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-a_{2,3} & -a_{1,3} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-a_{3,4} & -a_{2,4} & -a_{1,4} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -a_{3,5} & -a_{2,5} & -a_{1,5} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -a_{3,6} & -a_{2,6} & -a_{1,6} & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -a_{3,7} & -a_{2,7} & -a_{1,7} & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -a_{3,8} & -a_{2,8} & -a_{1,8} & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -a_{3,9} & -a_{2,9} & -a_{1,9} & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -a_{3,10} & -a_{2,10} & -a_{1,10} & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -a_{3,11} & -a_{2,11} & -a_{1,11} & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -a_{3,12} & -a_{2,12} & -a_{1,12} & 1
\end{pmatrix}$$

and

$$\Phi_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{3,1} & a_{2,1} & a_{1,1} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{3,2} & a_{2,2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{3,3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Appendix B

To fit a PAR(p), the first step is to select the order p. Here, we use AIC and hypothesis test on $a_{p+1,m} = 0$ to determine p. The results are reported in Table 8. When $p = 5$, AIC reaches the lowest value and the p_value of $a_{5,m} = 0$ is significant. p_value of $a_{6,m} = 0$ is not significant. Therefore, we will choose a PAR(5) with seasonal intercept for the time dependent model output.

We also test for the periodicity in the autoregressive parameters, and the results suggest that a PAR model fits better to the data.

```
Test for periodicity in the autoregressive parameters .
Null hypothesis: AR( 5 ) with the selected deterministic components.
Alternative hypothesis: PAR( 5 ) with the selected deterministic components.
```

Table 8.: Order selection of PAR(p)

Criterion	p				
	1	2	3	4	5
AIC	1494	1490	1500	1501	**1473**
p-value	0.1788	0.9573	0.6056	**0.0165**	0.7624

```
F-statistic: 1.47 on 55 and 158 DF, p-value: 0.0352 *
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Periodic integration test shows that there is no seasonal unit root. So, the model process is not PIAR.

```
> nsdiffs(nep.avg.ts, test="ch") # average trend
[1] 0
```

REFERENCES

REFERENCES

Veronica J. Berrocal, Peter F. Craigmile, and Peter Guttorp. Regional climate model assessment using statistical upscaling and downscaling techniques. *Environmetrics*, 23(5):482â€“492, 2012. ISSN 1099-095X.

M. Chen, Q. Zhuang, D. R. Cook, R. Coulter, M. Pekour, R. L. Scott, J. W. Munger, and K. Bible. Quantification of terrestrial ecosystem carbon dynamics in the conterminous united states combining a process-based biogeochemical model and MODIS and AmeriFlux data. *Biogeosciences*, 8(9):2665–2688, September 2011. ISSN 1726-4189. doi: 10.5194/bg-8-2665-2011.

Min Chen and Qianlai Zhuang. Spatially explicit parameterization of a terrestrial ecosystem model and its application to the quantification of carbon dynamics of forest ecosystems in the conterminous united states. *Earth Interactions*, 16(5):1–22, April 2012. ISSN 1087-3562.

Stefano Conti and Anthony O'Hagan. Bayesian emulation of complex multi-output and dynamic computer models. *Journal of Statistical Planning and Inference*, 140 (3):640–651, March 2010. ISSN 0378-3758.

Stefano Conti, Clive W Anderson, Marc C Kennedy, and Anthony O'Hagan. A bayesian analysis of complex dynamic computer models. In *Proc. of the 4th International Conference on Sensitivity Analysis of Model Output*, 2004.

Erwin Fehlberg. Low-order classical runge-kutta formulas with stepsize control and their application to some heat transfer problems. *NASA Technical Report 315*, 1969.

Philip Hans Franses. Periodicity and stochastic trends in economic time series. *OUP Catalogue*, 1996.

Philip Hans Franses and Richard Paap. Model selection in periodic autoregressions? *Oxford Bulletin of Economics and Statistics*, 56(4):421–439, 1994.

Dave Higdon, James Gattiker, Brian Williams, and Maria Rightley. Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, 103(482):570–583, June 2008.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3): 425–464, 2001. ISSN 1467-9868.

Danie G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society*, 52: 119–139, 1951.

Fei Liu and Mike West. A dynamic modelling strategy for bayesian computer model emulation. *Bayesian Analysis*, 4(2):393–411, June 2009. ISSN 1936-0975.

Kim L. Mueller, Vineet Yadav, Peter S. Curtis, Chris Vogel, and Anna M. Michalak. Attributing the variability of eddy-covariance co2 flux measurements across temporal scales using geostatistical regression for a mixed northern hardwood forest. *Global Biogeochemical Cycles*, 24(3), 2010. ISSN 1944-9224.

Jeremy Oakley and Anthony O'Hagan. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4):769–784, 2002.

Jeremy E Oakley and Anthony O'Hagan. Probabilistic sensitivity analysis of complex models: a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769, 2004.

Marcello Pagano. On periodic and multiple autoregressions. *The Annals of Statistics*, pages 1310–1317, 1978.

Abhijit Patil and Zhiqiang Deng. Temporal scale effect of loading data on instream nitrate-nitrogen load computation. *Water science and technology: a journal of the International Association on Water Pollution Research*, 66(1):36–44, 2012. ISSN 0273-1223.

Stephen D. Prince and Samuel N. Goward. Global primary production: A remote sensing approach. *Journal of Biogeography*, 22(4/5):815–835, July 1995. ISSN 0305-0270.

JW Raich and Wo H Schlesinger. The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate. *Tellus B*, 44(2):81–99, 1992.

Daniel B Rowe. *Multivariate Bayesian statistics: models for source separation and signal unmixing*. CRC Press, 2002.

Steven W. Running, Ramakrishna R. Nemani, Faith Ann Heinsch, Maosheng Zhao, Matt Reeves, and Hirofumi Hashimoto. A continuous satellite-derived measure of global terrestrial primary production. *BioScience*, 54(6):547–560, June 2004. ISSN 0006-3568.

Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

James H. Stapleton. *Linear Statistical Models*. John Wiley & Sons, 1995. ISBN 0-471-57150-4.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer, 1999.

Jinyun Tang and Qianlai Zhuang. A global sensitivity analysis and bayesian inference framework for improving the parameter estimation and prediction of a process-based terrestrial ecosystem model. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 114(D15), 2009.

Claudia Tebaldi, Richard L Smith, Doug Nychka, and Linda O Mearns. Quantifying uncertainty in projections of regional climate change: A bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 18(10):1524–1540, 2005.

David P. Turner, Michael Guzy, Michael A. Lefsky, William D. Ritts, Steve Van Tuyl, and Beverly E. Law. Monitoring forest carbon sequestration with remote sensing and carbon cycle modeling. *Environmental Management*, 33(4):457–466, August 2004. ISSN 0364-152X, 1432-1009.

Daniel Williamson, Michael Goldstein, Lesley Allison, Adam Blaker, Peter Challenor, Laura Jackson, and Kuniko Yamazaki. History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim Dyn*, 41(7-8):1703–1729, October 2013. ISSN 0930-7575, 1432-0894.

Jingfeng Xiao, Qianlai Zhuang, Beverly E. Lawc, Jiquan Chend, and Dennis D. Baldocchie. A continuous measure of gross primary productivity for the conterminous u.s. derived from modis and ameriflux data. *Remote sensing of environment*, 114: 576–591, 2010. ISSN 0034-4257.

Dongbin Xiu and George Em Karniadakis. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Computer Methods in Applied Mechanics and Engineering*, 191(43):4927–4948, 2002.

Feihua Yang, Kazuhito Ichii, Michael A. White, Hirofumi Hashimoto, Andrew R. Michaelis, Petr Votava, A-Xing Zhu, Alfredo Huete, Steven W. Running, and Ramakrishna R. Nemani. Developing a continental-scale measure of gross primary production by combining modis and ameriflux data through support vector machine approach. *Remote Sensing of Environment*, 110(1):109–122, September 2007. ISSN 0034-4257.

Hao Zhang. *Multivariate Geostatistical Models: Inference and Computing*. Chapman & Hall, London, 2014.

Li Zhang, Bruce Wylie, Thomas Loveland, Eugene Fosnight, Larry L. Tieszen, Lei Ji, and Tagir Gilmanov. Evaluation and comparison of gross primary production estimates for the northern great plains grasslands. *Remote Sensing of Environment*, 106(2):173–189, January 2007. ISSN 0034-4257.

Qing Zhu and Qianlai Zhuang. Improving the quantification of terrestrial ecosystem carbon dynamics over the united states using an adjoint method. *Ecosphere*, 4(10): art118, 2013.

Qing Zhu and Qianlai Zhuang. Parameterization and sensitivity analysis of a process-based terrestrial ecosystem model using adjoint method. *Journal of Advances in Modeling Earth Systems*, 6(2):315–331, 2014. ISSN 1942-2466.

Q Zhuang, AD McGuire, JM Melillo, JS Clein, RJ Dargaville, DW Kicklighter, RB Myneni, J Dong, VE Romanovsky, J Harden, et al. Carbon cycling in extratropical terrestrial ecosystems of the northern hemisphere during the 20th century: a modeling analysis of the influences of soil thermal dynamics. *Tellus B*, 55(3): 751–776, 2003.

Qianlai Zhuang, Tonglin Zhang, Jingfeng Xiao, and Tianxiang Luo. Quantification of net primary production of chinese forest ecosystems with spatial statistical approaches. *Mitigation and adaptation strategies for global change*, 14(1):85–99, 2009.

VITA

VITA

Xian He was born in Xiantao, Hubei, China. She obtained her Bachelor of Science degree in Applied Mathematics in 2006 from Huazhong University of Science and Technology. In 2009, she received her Master of Science degree in financial mathematics from Nankai University. After that, she moved to US and obtained a Master's degree in mathematical statistics at Purdue University in 2012.