Friedrich Schiller Universität Jena
Max-Planck-Institut für Biogeochemie Jena

Gaussian Process Regression for Uncertainty Estimation on Ecosystem Data

Diplomarbeit

zur Erlangung des akademischen Grades
Diplom-Bioinformatiker

vorgelegt von

Olaf Menzer

geboren am 20.03.1986 in Jena

Betreuerin: Antje Maria Moffat
Themenverantwortlicher: Prof. Ernst Günther Schukat-Talamazzini

August 5, 2011

Zusammenfassung

Der Kohlenstoffaustausch zwischen terrestrischen Ökosystemen und der Atmosphäre ist hauptsächlich gekennzeichnet durch nicht-lineare, komplexe und zeitversetzte Prozesse. Die damit verbundenen Ökosystemantworten und klimatischen Rückwirkungen zu verstehen, ist eine fundamentale Herausforderung im Hinblick auf Probleme wie steigende CO_2 Anteile in der Atmosphäre und den Klimawandel allgemein. Für gewöhnlich werden die zu Grunde liegenden Zusammenhänge als fest vorgeschriebene Funktionen modelliert, die eine Reihe von meteorologischen, Strahlungs- und Gasaustauschvariablen verknüpfen. Im Gegensatz dazu ermöglichen Algorithmen des Maschinellen Lernens (ML), wie zum Beispiel Künstliche Neuronale Netze oder Gaußsche Prozesse, das Erforschen der Zusammenhänge direkt auf Grundlage der Daten.

Mikrometeorologische, hochaufgelöste Messungen an Türmen, über den Globus verteilt, sind ein wichtiges Werkzeug für das Quantifizieren der Ökosystemvariablen: z.B. CO_2-Austausch, Sonnenstrahlung und Lufttemperatur werden fortlaufend gemessen. Um die Interaktionen und Rückwirkungen zwischen diesen Variablen besser untersuchen zu können, müssen mehrere schwierige Dateneigenschaften berücksichtigt werden: verrauscht, mehrdimensional und lückenhaft.

In dieser Arbeit wird das Abschätzen der Unsicherheiten in solchen mikrometeorologischen Messungen mit der Methode Gaußsche Prozesse (GPs) angegangen, einer modernen, nicht-parametrischen Methode für nicht-lineare Regression. In den letzten Jahren wurde gezeigt, dass die GP Methode ein mächtiger Modellierungsansatz ist, unabhängig von der Variablen-Dimensionalität, dem Grad der Nicht-Linearität und der Stärke des Rauschens. Heteroskedastische Gaußsche Prozesse (HGPs) sind eine GP Methode speziell für Daten mit einer variierenden, inhomogenen Rauschvarianz, wie sie gewöhnlich in Messungen des CO_2-Austausch der Fall ist. Hier zeigt eine Bewertung der HGP Leistung in mehreren Experimenten mit künstlichen Daten, sowie ein Vergleich zu existierenden Methoden, dass ihre besondere Fähigkeit darin liegt, unter relativ wenigen Voraussetzungen, das Rauschen in Messungen abzuschätzen und gleichzeitig gute Daten "Fits" zu liefern.

Auf der Grundlage von lückenhaften, verrauschten, halbstündlichen Messungen von Ökosystemvariablen werden HGPs eingesetzt um an zwei Messtürmen in Hainich (Deutschland) und Hesse (Frankreich), Unsicherheiten in Jahressummen des CO_2-Austauschs abzuschätzen. Ähnliche Rauschmuster mit verschiedenen Stärken wurden ermittelt, mit einem geschätzten jährlichen Rauschanteil von $\pm 14.1\ \mathrm{gCm^{-2}yr^{-1}}$ bzw. $\pm 23.5\ \mathrm{gCm^{-2}yr^{-1}}$ für das Jahr 2001.

Abstract

The flow of carbon between terrestrial ecosystems and the atmosphere is mainly driven by nonlinear, complex and time-lagged processes. Understanding the associated ecosystem responses and climatic feedbacks is a key challenge regarding climate change questions such as increasing atmospheric CO_2 levels. Usually, the underlying relationships are implemented in models as prescribed functions which interlink numerous meteorological, radiative and gas exchange variables. In contrast, supervised Machine Learning algorithms, such as Artificial Neural Networks or Gaussian Processes, allow for an insight into the relationships directly from a data perspective.

Micrometeorological, high resolution measurements at flux towers spread across the globe are an essential tool for obtaining quantifications of the ubiquitious ecosystem variables, as they continuously record e.g. CO_2 exchange, solar radiation and air temperature. In order to understand the interactions and feedbacks between these variables, several challenging data properties need to be taken into account: noisy, multidimensional and incomplete.

In this work, the task of investigating relationships and estimating uncertainties in such measurements was addressed by Gaussian Processes (GPs), a modern nonparametric method for nonlinear regression. The GP approach has recently been shown to be a powerful modeling tool, regardless of the input dimensionality, the degree of nonlinearity or the noise level. Heteroscedastic Gaussian Processes (HGPs) are a specialized GP method for data with a varying, inhomogeneous noise variance, as often observed in CO_2 flux measurements. Here, an evaluation of the HGP performance on several artificial experiments and comparison to existing nonlinear regression methods showed that their outstanding ability is to capture measurement noise levels, concurrently providing reasonable data fits under relatively few assumptions.

On the basis of incomplete, noisy half-hourly measured ecosystem data, HGPs were employed to assess uncertainties for annual sums of CO_2 exchange at the two flux tower sites in Hainich, Germany and Hesse, France. Similar noise patterns showing different magnitudes were detected, with annual random error estimates of ± 14.1 $gCm^{-2}yr^{-1}$ and ± 23.5 $gCm^{-2}yr^{-1}$, respectively, for the year 2001.

Contents

List of Figures

Abbreviations

ANN	Artificial Neural Network
CO_2	Carbon Dioxide
ELR	(Measured) Ecosystem Light Response Data
GP	Gaussian Process
GPML	Gaussian Process implemented with GPML Toolbox
HGP	Heteroscedastic Gaussian Process
LOWESS	Locally Weighted Scatterplot Smoothing
ML	Machine Learning
NLR	Least Squares Nonlinear Regression
SLR	Simulated Light Response Data

α	Initial Quantum Yield
ER_d	Daytime Ecosystem Respiration [μmol CO_2 m^{-2} s^{-1}]
GPP	Gross Primary Production [μmol CO_2 m^{-2} s^{-1}]
GPP_{opt}	Optimum GPP [μmol CO_2 m^{-2} s^{-1}]
NEP	Net Ecosystem Production [μmol CO_2 m^{-2} s^{-1}]
NEP^*_{sat}	NEP at maximum light intensity [μmol CO_2 m^{-2} s^{-1}]
$PPFD$	Photosynthetic Photon Flux Density [μmol photon m^{-2} s^{-1}]
$PPFD_{dif}$	Diffuse $PPFD$ [μmol photon m^{-2} s^{-1}]
$PPFD_{dir}$	Direct $PPFD$ [μmol photon m^{-2} s^{-1}]
R_{eco}	Ecosystem Respiration [μmol CO_2 m^{-2} s^{-1}]
Rb	Base Respiration at Reference Temperature [μmol CO_2 m^{-2} s^{-1}]
T_a	Air Temperature [°C]
$time_d$	Daily Binned half-hourly Timesteps [$days$]

$Bias$	Bias Error
$cov(y)$	$(= k(x, x'))$ Covariance Function of a Gaussian Process
$\mathbb{E}$	Expected Value
HBLR	Hyperbolic Light Response Curve
$\mathbf{I}$	The Identity Matrix
$\mathbf{K}$	Covariance Matrix of a Gaussian Process
$\mathbf{L}$	Cholesky Decomposition
$m(x)$	Mean Function of a Gaussian Process
$\mathcal{N}$	Normal Distribution, $\mathcal{N}(\mathbb{E}, \sigma^2)$
R^2	Coefficient of Determination
$RMSE$	Root Mean Square Error
σ_f	Signal Variance Hyperparameter of a Gaussian Process
σ_l	Length-scale Hyperparameter of a Gaussian Process
σ_n	Noise Hyperparameter of a Gaussian Process
f_{SS}	Smith Sigmoid Function
SSR	Sum of Squared Residual Function
σ	Standard Deviation
σ^2	Variance
σ_{GPML}	GPML predictive Standard Deviation
σ_{HGP}	HGP predictive Standard Deviation
$SDev$	Standard Deviation of the Laplacian Distribution
s	Estimated Standard Deviation (Linear Regression)
s^2	Estimated Variance (Linear Regression)
$se(\hat{y_0})$	Standard Error of the Expected Value at x_0 (Linear Regression)
τ_i	Confidence Iinterval for the Expected Value at x_i (Noninear Regression)

1 Introduction

This diploma thesis addresses the task of modeling the carbon exchange between terrestrial ecosystems and the atmosphere, defined as the ecosystem response to meteorological drivers. The underlying relationships are typically nonlinear, complex, time-lagged and involve autocorrelative effects (Moffat *et al.*, 2010). Micrometeorological, high resolution measurements at flux towers are essential for exploring the interactions and feedbacks of the processes regarding key climate change questions, such as increasing atmospheric CO_2 levels. Despite significant improvements in the applied instruments and measurement techniques during the last decades, there is still some random noise left in the measured ecosystem data, which it is challenging to assess. A specification of data uncertainty affects not only the uncertainty of the model, but also model predictions (Richardson *et al.*, 2006):

"From the standpoint of model-data synthesis, uncertainties are as important as the data values themselves" (Raupach *et al.*, 2005)

The motivating background of modeling biosphere-atmosphere interactions and estimating relevant uncertainties in the ecosystem measurements is to better the understanding of the global carbon cycle, a subject which has long attracted scientists from various disciplines. Rising concentrations of greenhouse gases in the atmosphere have been attributed to industrialization, human development and the resulting combustion of fossil fuels and changes in land use. The concentration of carbon dioxide (CO_2), the most important green house gas after water vapour, is currently the highest it has been in the last 650.000 years (Siegenthaler *et al.*, 2005). The terrestrial biosphere strongly influences the global carbon cycle by sequestering carbon via photosynthesis while simulatenously releasing carbon via respiration. It was found that it exchanges 123 ± 8 Gt of carbon per year with the atmosphere (Beer *et al.*, 2010), while there are also considerable exchanges between oceans and atmosphere ($\sim$90 Gt of carbon, see Fig. 1.1). Terrestrial Ecosystems are often viewed as long-term carbon sinks, although their carbon seques-

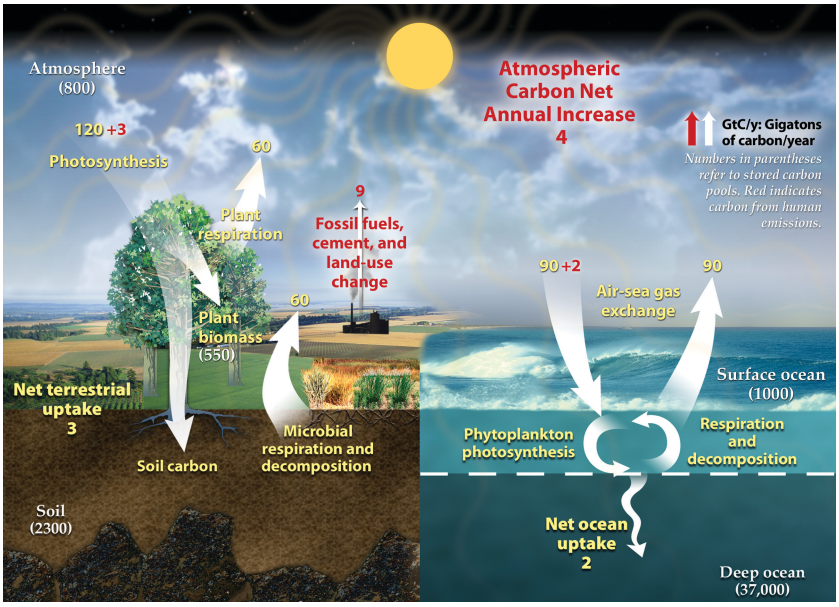


Figure 1.1: The global carbon cycle, illustrating the movement of carbon between land, atmosphere, and oceans. Yellow numbers are natural fluxes, and red numbers are human contributions in gigatons (Gt) of carbon per year. White numbers indicate stored carbon (U.S. Department of Energy, 2008).

tration capacity can vary largely between e.g. temperate deciduous forests and boreal coniferous forests (Valentini *et al.*, 2000; Baldocchi *et al.*, 2001). The terrestrial carbon sink is also considerably influenced by CO_2 fertilization, climate change, stand age and recovery from disturbance (Schulze *et al.*, 2000).

Providing high resolution measurements of the net CO_2, H_2O and energy fluxes, as well as auxiliary meteorological variables, over a wide range of ecosystems, the FLUXNET observational network (Baldocchi, 2008) is an essential data source towards identifying the contributions of various ecosystems to a global carbon sink. The CO_2 exchange between biosphere and atmosphere is measured as the Net Ecosystem Production (NEP), which equals the difference between the carbon assimilation by photosynthesis (gross primary production, GPP) and the release of carbon to the atmosphere (ecosystem respiration, ER). The obtained datasets measured by the eddy covariance technique have the following properties: complex, noisy, multidimensional and fragmented (Moffat, Accepted).

Usually, the underlying ecosystem responses are implemented in models as prescribed functional relationships. In contrast, supervised Machine Learning (ML) algorithms, such as Artificial Neural Networks, allow to extract the relationships to be characterized directly from the data (Moffat, Accepted). The term *Machine Learning* refers to

intelligent methods, by which systems or computers learn characteristics and patterns through generalizing from sample data. These tasks include prediction (both regression and classification), planning and robot control. One of the modern ML methods, Gaussian Processes (GPs), were shown to be a powerful tool for nonlinear regression, regardless of the input dimensionality, the degree of nonlinearity or the noise level (Rasmussen, 1996) and therefore are a promising method to be applied on ecosystem data with the above properties. More recently, they have been shown to be applicable to real world problems in biological or financial models (Gao, 2004; Sun *et al.*, 2010; Macke *et al.*, 2011). Being a probabilistic model, GPs are also known for their good ability to estimate uncertainties directly from the posterior distribution and are therefore the method of choice.

Here, the central focus is on assessing the uncertainty, both, in the FLUXNET measurements, as well as, the extracted relationships between fluxes and meteorological drivers, by evaluating prediction and confidence intervals, respectively. Uncertainty estimates in an ecophysiological context are of importance because they allow to quantify the mismatch between models and data and can thus be used for model optimization, ecosystem model validation against flux data, multi-site syntheses or regional-to-continental integration efforts (Raupach *et al.*, 2005; Richardson *et al.*, 2008). Consequently, an understanding of the uncertainties is crucial for the use of the FLUXNET observations to constrain future climate predictions. Hence, from a data perspective, the GP method promises to be a suitable approach to study.

This work is an attempt to present GPs as a novel method to explore ecological data sets, especially regarding uncertainties in measurements. First, their ability to model nonlinear relationships and to estimate uncertainties is tested on simulated data, where the expected outcome is known beforehand. Afterwards they are applied to real world data using their particular strengths, either as a stand-alone or an accompanying method. A comparison to least squares nonlinear regression (NLR) and other methods, such as local weighted regression smoothing (LOWESS), will give further insights into the performance of GPs.

To demonstrate the principles of the GP method, it is first portrayed from a theoretical perspective and compared to classical regression methods (Chapter 2). The main part of this work, a series of artificial and real world data experiments, is then presented and evaluated in order to point out their particular strengths and applicability (Chapter 3). The last chapter comments on subsequent conclusions and a future perspective.

2 Materials and Methods

In any system in which variable quantities change, it is of interest to examine the effects that some variables exert (or appear to exert) on others (Draper & Smith, 1998), this is the general understanding of the term *relationship* in data analysis. The statistical field of regression analysis comprises a collection of so-called *learning methods* to quantify, (mathematically) describe and understand that relationship between a target variable (or dependent variable) and one or more input variables (or independent variables). With a regression analysis it is also possible to make *predictions* about some unmeasured events. Usually, learning methods are trained to serve one of the two above goals.
The regression analysis is a technique which is widely employed for the understanding of numerous real-world problems in the life sciences and environmental sciences, e.g. in the fields of gene expression analysis (Müller *et al.*, 2008), enzyme kinetics (Duggleby, 1995), plant physiology (Storch & Zwiers, 1999; Reichstein *et al.*, 2005) or climatology (Drignei *et al.*, 2008). The principle of "learning from examples" has not only attracted biologists and psychologists interested in the interaction of organisms with their environment, but also mathematicians and computer scientists who are mainly studying learning in artificial contexts.
This chapter introduces the data domain, including data acquisition methods (Section 2.1.1), a measurement site description (Section 2.1.2) and the according ecosystem data properties (Section 2.1.3) including a review on the current knowledge about flux data uncertainties. The following sections focus on the method of regression analysis. In Section 2.2, linear regression, which is used when a linear relationship between the target variable and the input varaiables can be assumed. In the situation of a nonlinear relationship, as often observed in natural science problems, there is a variety of applicable different methods and approaches ranging from statistical approaches (e.g., nonlinear least squares, Section 2.3), over Artificial Neural Networks to newer non-parametric models such as Gaussian Processes and other Machine Learning methods (e.g., Support Vector Machines). Gaussian Processes are a probabilistic model used for nonlinear regression, being the key method of this thesis they are separately described in Section

2.4.

2.1 Data Domain

2.1.1 Eddy Covariance Method

The eddy covariance technique is a micrometeorological method for measuring the turbulent fluxes in the atmospheric boundary layer, thus making it possible to quantify the turbulent exchange of energy and matter between surface and atmosphere. The word *eddy* refers to the turbulent fluxes (short for flux densities) in the air (Fig. 2.1). It was first published in the mid 20$^{\text{th}}$ century by Montgomery (1948), and then benefited a lot from the development of the sonic anemometer (Bovscheverov & Voronov, 1960) which enabled high temporal resolution measurements.

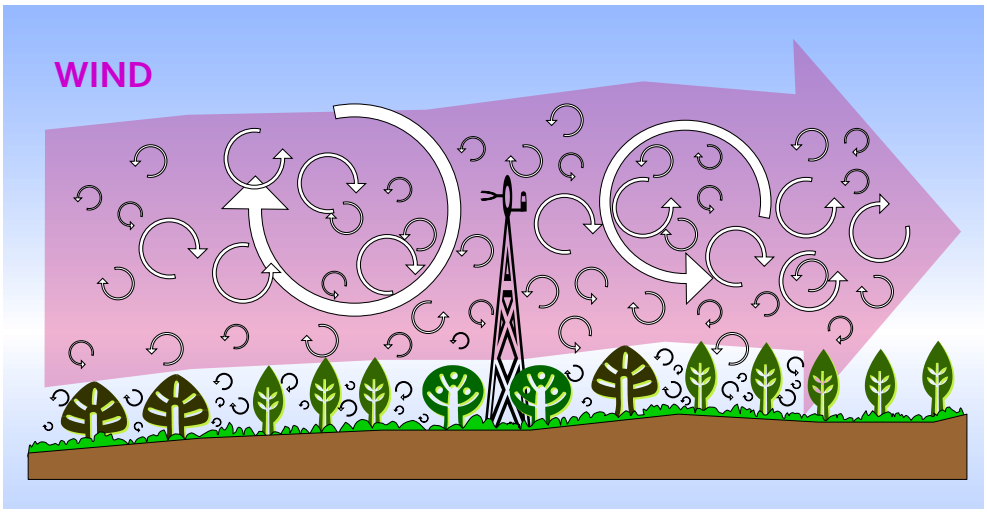


Figure 2.1: Sketch of the turbulent ecosystem fluxes (*eddies*) next to a measurement site tower (Burba & Anderson, 2010).

The fluxes (of e.g., CO_2) can be described by the following equation (Foken, 2003)

$$F_C = \rho_a \overline{w'c'}, \tag{2.1}$$

where ρ_a is the molar density of the air, w is the vertical wind speed and c the substrate concentration. Overbars denote temporal averages and primes denote short term deviations from the mean. Roughly spoken, this equation states that the net flux can be calculated from the time averaged covariance between the substrate concentration and

the vertical wind vector. To determine these covariances, high frequency measurements of the wind vector are taken using three-dimensional sonic anemometers (Moncrieff *et al.*, 1996). Before calculating the mean net flux densities by averaging usually over half-hourly timespans, various data correction steps, such as high-frequency losses, are necessary (Aubinet *et al.*, 2000).
The eddy covariance method is based on relatively few assumptions such as turbulent conditions, horizontal homogeneity in the vegetation and a flat terrain. Another advantage is that in assessing the carbon flux, eddy covariance measurements are scale appropriate and survey a whole ecosystem, whereas cuvette and chamber systems only capture a small part of it (Baldocchi, 2003).
It is important to mention that the so-called *footprint*, the source area of the eddy fluxes, can vary temporally and is also depending on the installation height of instruments at the flux tower. This height can be fairly different for e.g. grasslands and forests, resulting in typical footprint ranges between 100 m to 2000 m (Schmid, 1994). Recently, state-of-the-art footprint models (Goeckede *et al.*, 2006) have been developed also for sites in more complex, horizontal heterogeneous terrains.

2.1.2 The Hainich Flux Tower Site

One of the measurement sites of ecosystem flux data is the tower in the "Hainich National Park" in Central Germany (51°04'46" N, 10°27'08" E, 440 m a.s.l.). The forest at this experimental site was unmanaged for more than 60 years prior to 1997 because of its history as a military base. Also, in the centuries before, the area was used as a coppice with standard-systems and therefore not exposed to clearcut. Hence, the Hainich forest developed basically undisturbed and the trees cover a wide range of age classes with a maximum of up to 250 years (Knohl, 2003). The dominating species in this deciduous forest are beech (*Fagus sylvatica*, 65%), ash (*Fraxinus excelsior*, 25%) and maple (*Acer pseudoplantanus* and *Acer plantanoides*, 7%).
In the footprint of the tower, the stand density is 334 trees/ha, the typical canopy height is 33 m the and there are trees with a maximum height of 37 m. In 2010, a completely new flux tower was set up (Fig. 2.2), which has instruments installed at a measurement height of 45 m, whereas the measurement height of the old tower was at 43.5 m.
The climate at the tower site is suboceanic/subcontinental, with a long term annual mean temperature of $7.5 - 8$°C and annual precipitation of 750-800 mm. Please refer to e.g. Kutsch *et al.* (2008) or Knohl (2003) for a more detailed site and forest stand characteristics description.

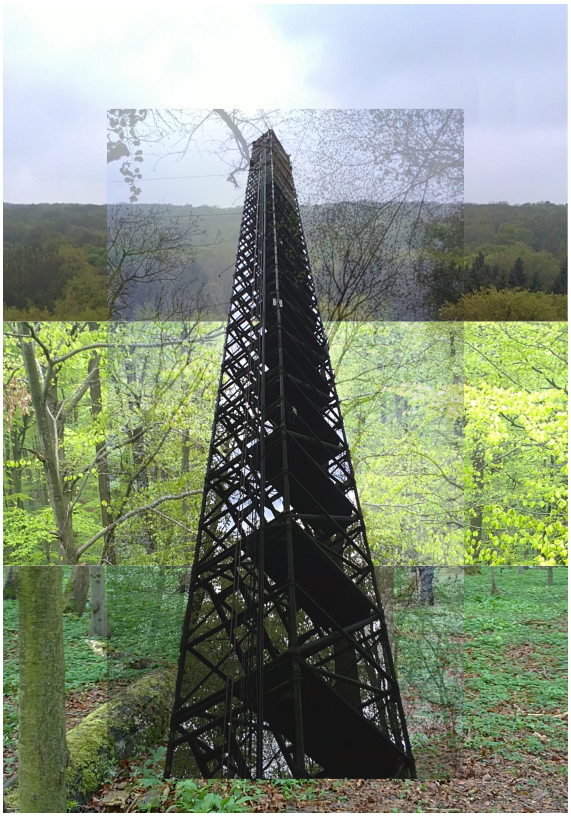


Figure 2.2: Montage of a photograph of the new Hainich flux tower (taken in April 2011), portrayed on the surrounding biosphere and atmosphere.

2.1.3 Ecosystem Data Uncertainties

The measured ecosystem fluxes have various properties which are comprehensively characterized in Moffat (Accepted). For this study, the most relevant of them is the random error in the measurements. Important other properties shall only be summarized here briefly:

- **Incompleteness:** Due to limitations of the eddy covariance technique (e.g., instrumentation failure, footprint issues or horizontal advection flow), a considerable percentage (20 - 60%) of the annual carbon flux measurements needs to be rejected. Most of these so-called *gaps* occur during nighttime because of non-turbulent conditions. There are various approaches in the literature for filling these gaps (Falge *et al.*, 2001; Moffat *et al.*, 2007).

- **Multidimensionality:** There are numerous other meteorological and ancillary variables measured, each of them with its own measurement noise and occasional

gaps.

- **Inconsistency:** Even for measurements under very similar meteorological conditions, the measured carbon flux can be different due to changes in the state of the ecosystem (e.g., phenology, soil properties or time lag effects).

In general, the sources of error in the flux data need to be distinguished. On the one hand, there are *random errors*, which are due to measurement instruments, stochastic nature of turbulences and the varying footprint of the towers. On the other hand, flux data is subject to *systematic errors*, which are caused by inaccurate calibration, measurements under unfavorable meteorological conditions (e.g., at night) or issues related to advection and non-flat terrain. The systematic errors are usually difficult to detect or quantify in data-based analysis (Moncrieff *et al.*, 1996). According to Hollinger & Richardson (2005), the uncertainties in the flux data are largely due to random measurement error. However, if one is to completely describe the *total* flux measurement error, it also requires a quantification of the systematic error or bias (Goulden *et al.*, 1996).
There are various ways of assessing these uncertainties to find out how much noise is in the measurements. The most recent findings are listed below:

0) Paired tower method (Hollinger *et al.*, 2004): A second tower in a spruce dominated forest in Maine, USA, was set up very close to an existing flux tower, with their footprints not overlapping significantly. Thus, the meteorological conditions including temperature and radiation were almost identical for both towers.
The between tower variability was found to be lower than the interannual variability in NEP. The standard deviation σ of the flux data uncertainties could therefore be calculated by estimating the standard deviations of the differences between the two towers, which correspond to the expected error magnitude. Relevant findings suggest that CO_2 flux uncertainty varies with the season and declines with increasing wind speed. (Note: this method is enumerated as "0." because it is rather an exceptional study and of course not applicable at every site)

1) Paired observation method (Richardson *et al.*, 2006): This approach is analogue to 0), just with "time traded for space", i.e. flux measurements are compared on two successive days at exactly the same time of the day. Nearly identical environmental conditions were assured by fixed thresholds of meteorological variables such as air temperature and wind speed. Since these constraints are frequently not met, sample sizes are smaller than in the previous method.

The results on σ are investigated for seven different (exclusively North American) measurement sites and generally in agreement with 0). A detailed analysis of the statistical properties of the flux uncertainties is also carried out in this study, i.e. the first four moments of the error distribution are calculated. The most relevant reported results are that the random error follows rather a double-exponential (Laplace) than a normal (Gaussian) distribution and it increases as a linear function of the flux magnitude.

2) Model residuals method (Richardson *et al.*, 2008): In the subsequent study to the paired observation approach, again the statistical properties in terms of the first four moments of the error distributions were investigated. The difference to the previous study is that these properties were inferred from the model residuals (the difference between model predictions and measured fluxes) of five different modeling approaches (e.g., NLR and ANN). The former result of an Laplacian error distribution that varies with the flux magnitude (i.e., a heteroscedastic error) could be confirmed. Also, a spectral analysis of the model predictions suggests autocorrelated model residuals, which exhibit site-specific differences. This study considered six European measurement sites, including the Hainich forest (Section 2.1.2).

3) Gap-filling algorithm method (Lasslop *et al.*, 2008): In this approach, the random errors are estimated using the gap-filling algorithm of Reichstein *et al.* (2005). The result of heteroscedastic flux data errors that increase with the flux magnitude could be confirmed again. However, this study suggests that the error distribution is rather a superposition of Gaussian distributions than a Laplacian distribution. It is recommended by the authors to characterize the normalized error distribution, i.e. the error distribution scaled to unity, when characterizing uncertainties. Regarding the autocorrelation of the errors, it was found that it is usually below 0.6 at a lag of 0.5 h.

In contrast to the three previous methods, this approach also investigated the influence of systematic errors on parameter and uncertainty estimates. It is suggested that uncertainties in flux data are underestimated with approaches that neglect so-called *selective* systematic errors, that occur only under certain conditions (e.g., under unfavorable meteorological conditions).

2.2 Linear Regression

A linear relationship

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{2.2}$$

between a target variable Y and an input variable X, determined by the parameters β_0 and β_1 and accounting for some unobservable random noise ϵ, can be modeled by a linear regression model of the form

$$\hat{Y} = b_0 + b_1 X. \tag{2.3}$$

Note that the word "'linear"' in linear regression refers to a model which is linear in the parameters, but still can be nonlinear in the input variable X (e.g. $\hat{Y} = b_0 + b_1 X + b_2 X^2$). However, for notational simplicity, from here on the linear regression model is assumed to be the standard first-order model with two parameters β_0 and β_1.
The linear regression model corresponds to fitting a regression line through a set of observations given by $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, with $\hat{y_i}$ being the predicted value of y_i for a given x_i. The regression coefficients b_1 (the slope of the regression line) and b_0 (the intercept with the y-axis) denote the estimates of the true model parameters β_1 and β_0, which are unknown since the data only represents a finite subset of the truth. A model residual e_i at a location $i \in \{1, 2, ..., n\}$ is defined as the difference between the observed value and the value predicted by the model:

$$e_i = y_i - \hat{y}_i = y_i - b_0 + b_1 x_i. \tag{2.4}$$

The residuals quantify the amount by which every individual y_i falls off the regression line and are therefore estimates of the true, unobservable error ϵ_i.
There are several different methods for estimating the regression coefficients with linear regression such as Ordinary Least Squares, Generalized Least Squares or Maximum Likelihood estimation. The Ordinary Least Squares (OLS) approach will be explained in the following, in order to exemplify the principle of linear regression.
In the OLS method, the measure that is to be minimized is the so-called sum of squared residual (SSR) which is the sum of the squared model residuals:

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum (y_i - b_0 - b_1 x_i)^2. \tag{2.5}$$

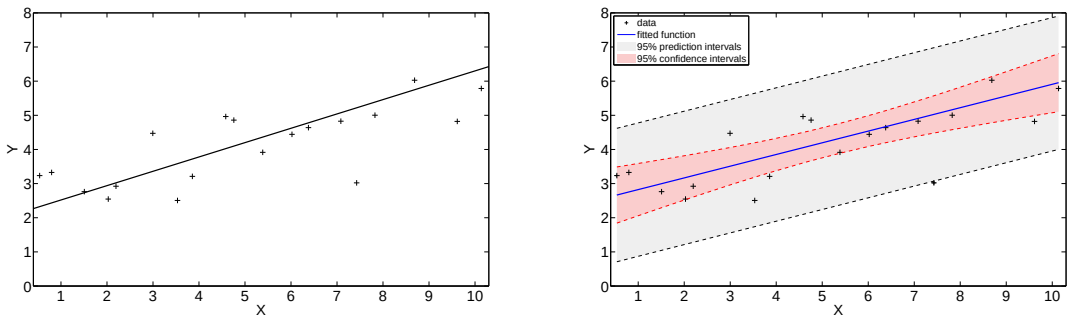
Now differentiating the SSR function first with respect to b_0 and then with respect to b_1 results in two equations by which the regression coefficients can be determined easily (for technical details please refer to Freedman (2009) or Draper & Smith (1998)) and denoted as:

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{(\sum x_i - \bar{X})^2}, \tag{2.6}$$

and

$$b_0 = \bar{Y} - b_1\bar{X}. \tag{2.7}$$

with $\bar{X}$ and $\bar{Y}$ representing the means of X and Y, respectively.
To demonstrate how the linear regression works in practice one can apply the equations above to a data sample where the generating function and the noise magnitude is known, e.g. 20 samples drawn from the linear function $f(X) = 2.1 + 0.42X + \epsilon$, with disturbance term ϵ being modeled as i.i.d. (identically, independently distributed) Gaussian noise (Fig. 2.3(a)). Gaussian noise means that the noise follows a normal distribution. Besides these properties of the errors and of course a linearity of the underlying relationship, another important assumption for linear regression is that the variance of the errors is homogeneous over the whole range of observations (homoscedasticity). These assumptions should always be checked by an analysis of residuals, which gives insight into the distribution of the deviations. The regression coefficients calculated with OLS are $b_0 = 2.48$ and $b_1 = 0.34$, and correspond to the regression line (blue line in Fig. 2.3(b))



(a) data sample (b) linear regression

Figure 2.3: Linear regression example. In panel (a) the generating function $f(X) = 2.1 + 0.42X + \epsilon$ and the data sample are shown, in panel (b) the linear regression fit (blue line) and the corresponding prediction intervals (gray shaded areas) and confidence intervals (red shaded areas) are depicted.

which best fits that specific data sample. The reason for the deviation of the regression coefficients from the true parameters is that with OLS one obtains the regression parameters which minimize the sum of the residuals of the data points to the regression line (in theory that sum is zero). Thus, especially for smaller samples (e.g., $n < 100$), the regression line differs to some extent from the true generating function due to the noise in the data. But since the true generating function is more of interest, it is necessary to provide confidence intervals for the fitted function. One needs to distinguish *confidence intervals* and *prediction intervals*, the latter can be derived from

$$s^2 = \frac{1}{n-p} \sum e_i^2, \tag{2.8}$$

with p referring to the degrees of freedom, e.g. in the simple first order model discussed above $p = 2$. s^2 is an estimate of the true (unknown) variance of the data, σ^2. The square root σ of this value is commonly known as the standard deviation. For normally distributed data, the 1.96-fold of the estimated error variance s^2 corresponds to the so-called prediction intervals, i.e. the limits in which 95% of predicted data points are to be expected. The constant 1.96 approximately refers to the bounds by which 95% of the area under a normal curve are given. The prediction intervals are depicted for the regression in Fig. 2.3(b) by the gray shaded areas, whereas the red shaded areas show the confidence intervals for the function predicted by the linear regression. These serve a different purpose than the prediction intervals, i.e. they provide uncertainty estimates of the expected value of a given x_i. While s^2 is assumed to be constant along the x-axis for homoscedastic noise, the confidence intervals are varying and need to be estimated separately at every x_0 of interest. This can be done by calculating the standard error se as pointed out in Bates & Watts (1988):

$$se(\hat{y}_0) = s \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum (x_i - \bar{X})^2} \right)^{1/2}. \tag{2.9}$$

If a normal distribution can be assumed, 1.96 standard errors correspond to 95%-confidence intervals for the predicted linear regression line. Intuitively, the confidence interval has a minimum when $x_0 = \bar{X}$ and increases as x_0 is moved away from the mean. The prediction and confidence intervals are the key tools in this thesis for estimating uncertainties in data and relationships, and if technically possible they will always be provided for any of the regression methods used.

To quantify the general fit performance of a regression model, there are some stan-

dard statistical measures which are used throughout the whole thesis. The coefficient of determination (e.g., in Draper & Smith (1998))

$$R^2 = \frac{\{\sum (\hat{y}_i - \bar{Y})(y_i - \bar{Y})\}^2}{\sum (\hat{y}_i - \bar{Y})^2 \sum (y_i - \overline{Y})^2}, \tag{2.10}$$

describes the correlation between the modeled values and the measured values in terms of how well the regression line fits the given data. It is a fixed value in the range [0,1]; zero indicates that there is no correlation at all, whereas a regression line with an R^2 of 1 perfectly fits the data. Note that this is only one equation for the R^2 out of several and it should be used under the assumption of a linear relationship between the modeled and the observed data points.
The root mean square error ($RMSE$) is defined as

$$RMSE = \sqrt{\frac{1}{n} \sum (\hat{y}_i - y_i)^2}. \tag{2.11}$$

and serves as a single error measure of the model residuals of all (predicted) data points. It also serves as a statistic for the predictive power of the model and penalizes large prediction errors more than small prediction errors.
The mean bias error given by

$$Bias = \frac{1}{n} \sum (\hat{y}_i - y_i), \tag{2.12}$$

i.e. the mean deviation of the estimated values from the observed values, if it is zero than the estimator is said to be unbiased.
The standard deviation

$$SDev = \sqrt{(2)} \cdot \frac{1}{n} \sum |\hat{y}_i - y_i|, \tag{2.13}$$

of a Laplacian distribution is used in two the reference studies (Richardson *et al.*, 2006; Moffat, Accepted) for model residuals of ecosystem data.

2.3 Nonlinear Regression

Nonlinear regression models describe a relationship in which at least one of the parameters interacts with the other parameters in a nonlinear way. This Section concentrates on statistical methods to deal with nonlinear regression problems, such as the classical approach of nonlinear least squares (Section 2.3.1) and a more modern technique yet making use of the least squares principle, the local linear regression, better known as LOWESS (Section 2.3.2).
In nonlinear regression analysis, the model is a function that is a nonlinear combination of the parameters, represented by e.g. exponential functions, logarithmic functions or power functions. In fact, the possibilities of modeling a nonlinear relationship with such functions are endless. Thus, nonlinear regression is often applied in situations where definite information about the form of the relationship is available. This can include direct knowledge of the true model or might involve differential equations that the model must satisfy.

2.3.1 Least Squares Nonlinear Regression (NLR)

A nonlinear relationship

$$Y = f(X, \theta) + \epsilon, \tag{2.14}$$

between a target variable Y and an input variable X is determined by parameters $\theta = [\theta_1, ..., \theta_p]$ that interact in a nonlinear manner. Typical examples for such a relationship are

$$Y = \theta_1 X^{\theta_2} + \epsilon, \tag{2.15}$$

and

$$Y = \frac{\theta_1}{\theta_2} \cdot \exp(-\theta_2 X) + \epsilon. \tag{2.16}$$

Note that the first model can be transformed by simply taking logarithms to the base e into the form

$$\ln\ Y = \ln \theta_1 + \theta_2 \ln X + \epsilon, \tag{2.17}$$

which is linear in the parameters. Draper & Smith (1998) would call that model "intrinsically linear" since it is a nonlinear model that can be transformed into a linear form and is thus easier to handle. After transformation into the linear model the least squares method for linear regression (Section 2.2) can be applied. Note that after retransformation into

$$Y = \theta_1 X^{\theta_2} \cdot \exp(\epsilon), \tag{2.18}$$

the errors are then not additive anymore but multiplicative, also, they are lognormal distributed. Hence, a transformation should only be applied, when the assumptions about the errors are still (or even better) fulfilled by the linearized model. The distribution of the errors should be checked with an analysis of residuals.
The second example (eq. 2.16) cannot be transformed into a linear model, then the model is called "intrinsically nonlinear". Nevertheless, a transformation could still be applied here to make the fitting more easy.
In classical statistical analysis, the nonlinear regression is performed usually by a direct minimization of the sum of squared residual (eq. 2.5), which is not as easy as in the linear case. Differentiating the SSR function with respect to the parameters results in equations that are not linear and difficult to solve. Thus, estimating the parameters of a nonlinear regression fit requires heavy iterative calculations. Another problem is that the optimization algorithms applied sometimes do not find the optimal solution because of local minima in the parameter search space or because they oscillate around the optimum. Also, for some methods, it is of importance to provide initial parameter guesses that are not too far from the desired $\hat{\theta}$.
There are several established iterative techniques to minimize the SSR function of a nonlinear model, any of them requires such intensive calculations that can only be done computationally. Three examples shall be outlined here, without going into technical details. First, the Gauss-Newton algorithm approximates a linear expansion of the sum of squares function and then uses the results of linear least squares in a succession of stages. It is a relatively fast and simple method, but in some cases it might not converge, especially when the initial parameters are too far from the optimum.
The method of steepest descent is based on the idea to approach the minimum of the sum of squares function following a zig-zag path, where a new search direction is orthogonal to the previous one. The information, where the path through the parameter search space decreases most quickly is obtained by evaluating the partial derivatives with respect to the parameters. This method is robust when starting at a bad initial

guess, but may oscillate around an optimum without converging.
The so-called Levenberg-Marquardt algorithm takes advantage of both the strengths of Gauss-Newton and steepest descent and finds a compromise between them avoiding their most serious limitations. It is based on the work of Levenberg (1944) and Marquardt (1963) and probably the most widely applied algorithm for nonlinear estimation. In fact, when the current solution is far from the optimum, the algorithm behaves like steepest descent, whereas being close to the solution it becomes a Gauss-Newton method. The Levenberg-Marquardt algorithm directly makes use of the Jacobian matrix of the first partial derivatives with respect to p model parameters

$$\mathbf{J} = \begin{bmatrix} \frac{\partial f(x_1,\theta)}{\partial \theta_1} & \dots & \frac{\partial f(x_1,\theta)}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(x_n,\theta)}{\partial \theta_1} & \dots & \frac{\partial f(x_n,\theta)}{\partial \theta_p} \end{bmatrix}. \tag{2.19}$$

The Jacobian Matrix also serves for approximating the confidence and prediction intervals for a nonlinear estimation, which is different than in the linear case. In general, in the nonlinear case, the confidence intervals are more flexible and do not narrow towards the "centre of gravity" of the data. The confidence interval for a nonlinear fitted function can be denoted as (Bates & Watts, 1988)

$$\tau_i = s\sqrt{a_i^T(\mathbf{J}^T\mathbf{J})^{-1}a_i}, \tag{2.20}$$

with a_i being the i-th row of the Jacobian and s being the mean squared error (eq. 2.8). The prediction intervals can then be derived by

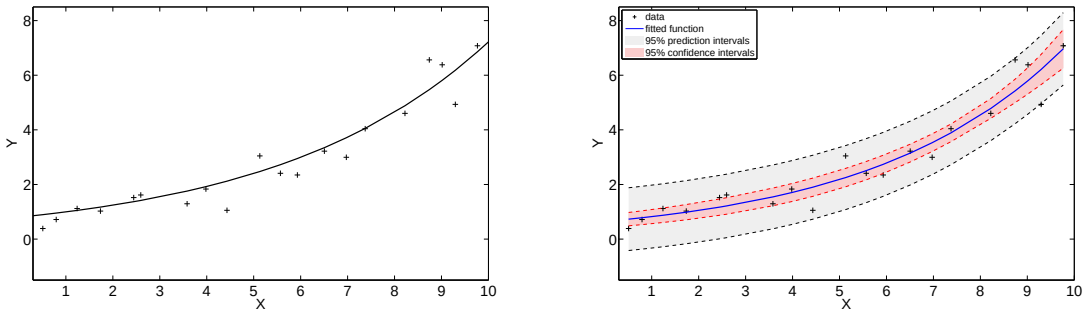
$$\hat{\sigma}_i = \sqrt{s^2 + \tau_i^2}. \tag{2.21}$$

Note that the prediction intervals are not constant anymore as in the linear regression on normal distributed data, but can differ at every x.
For illustration of the nonlinear least squares, the nonlinear model

$$Y = \theta_1 \cdot \exp(\theta_2 X) + \epsilon, \tag{2.22}$$

with parameters $\theta_1 = 0.8$ and $\theta_2 = 0.22$ was chosen and 20 data points were sampled, assuming i.i.d. noise terms again (Fig. 2.4(a)).
The Levenberg-Marquardt method resulted in the fit shown in Fig. 2.4(b) with the parameter estimates $b_1 = 0.65$ and $b_2 = 0.24$. The deviations of the fitted parameters

from the generating are, like in the example in Section 2.2, explainable by the noise in the data and the sample size. Note that for a nonlinear regression, apart from the above described confidence and prediction intervals, also confidence intervals for the parameters can be calculated, which will be omitted here for shortness.



(a) data sample (b) nonlinear regression

Figure 2.4: Nonlinear regression example. In panel (a) the generating function $f(X) = 0.8 \cdot \exp(0.24X) + \epsilon$ and the data sample are shown, in panel (b) the nonlinear regression fit (blue line) and the corresponding prediction intervals (gray shaded areas) and confidence intervals (red shaded areas) are depicted.

2.3.2 Locally Weighted Scatterplot Smoothing (LOWESS)

An example for a non-parametric regression model is the local regression smoothing which is referred to as LOESS or LOWESS (Locally Weighted Scatterplot Smoothing, Cleveland (1979)). The idea is to fit a nonlinear curve by stepwise local linear regressions. LOWESS, the method applied as a benchmark in this thesis, uses a linear polynomial model for the fit, whereas LOESS uses a quadratic polynomial model. In some literature, the term "LOESS" is used to refer to both of the smoothing methods. LOWESS combines linear and nonlinear regression by performing separate linear regressions at every x in a pre-defined span l, which is given as a fraction. Only the data points in l, i.e. the $(l \cdot N)$ nearest neighbours of x are used for the local weighted linear regression. The regression weights w_i are calculated for each data point in the span by a tri-cube weight function

$$w_i = \left(1 - \left|\frac{x - x_i}{d(x)}\right|^3\right)^3. \tag{2.23}$$

x is the predictor value associated with the response value to be smoothed, x_i are the nearest neighbors of x and $d(x)$ is the distance along the x-axis from x to the most distant predictor value within the span. It follows that the data point to be smoothed has the largest weight and thus the most influence on the local fit, whereas data points which are further away have smaller weights (Fig. 2.5(b)). Data points outside the span are set to zero weight and have no influence on the fit.

Recalling the data sample from the previous Section, the principle how LOWESS works is demonstrated for one data point x, marked in red in Fig. 2.5. A weighted linear least squares regression line is fitted through the according nearest neighbours of x, as depicted in Fig. 2.5(a). This procedure is repeated at every data point of the sample, resulting in a set of smoothed points, which can then be interpolated to a line that fits the data. The parameter l corresponds to the smoothness of the fit, for a larger l (e.g., $l = 0.8$) more data points are considered at every local regression step resulting in a smoother fitted curve. A smaller l makes the curve more sensitive to local function properties and to wiggle more strongly, as shown in Fig. 2.6(a). Often the smoothing parameter is set to values between 0.25 and 0.5, Cleveland (1979) suggested $l = 0.5$ as a reasonable starting value when there is no idea what is needed. There are also iterative approaches such as the PRESS procedure (Allen, 1974) for optimizing l.

Assumptions for LOWESS are that the errors in the data are independent and normally distributed. Also it should be checked that the fitted curve follows the pattern of the data, i.e. produces a nearly unbiased estimate.

A big advantage of LOWESS is that no function or model needs to be specified previously in order to fit a given data sample. Only the smoothing parameter and the degree of

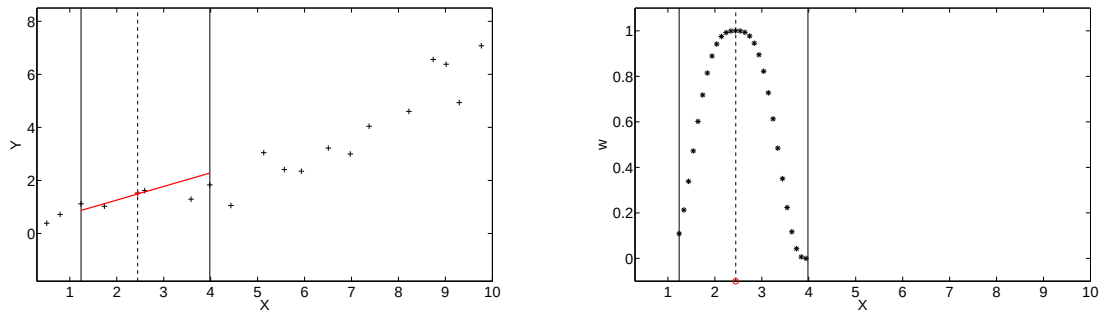
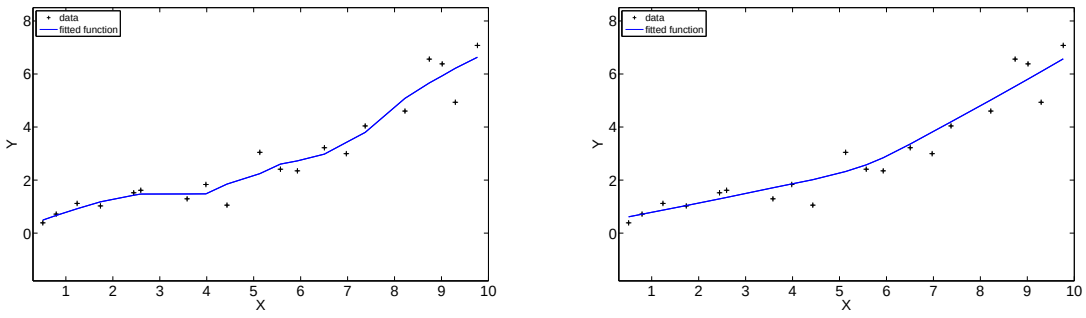


Figure 2.5: The same data sample as in Fig. 2.4. (a) Local regression span for the data point marked in red. The span parameter is set to $l = 0.3$, corresponding to 6 nearest neighbours. The red line depicts the local linear regression. (b) The according weights along the data span.

the local polynomial must be considered for optimization. Also, LOWESS shares the benefits of linear least squares regression such as given uncertainties for prediction and calibration. LOWESS is a flexible nonlinear regression method, which can be applied relatively easy for modeling complex relationships whose structure is rather unknown. However, to produce good models, LOWESS needs densely sampled data that forms a good empirical base for the fitted curve. The curve itself cannot be represented by a mathematical formula, which is another drawback of the model, especially when it comes to interpolation of incomplete data, although prediction with LOWESS is generally possible. LOWESS is also vulnerable to effects of outliers, although Cleveland (1979) provided some robust versions of the method. Another drawback of LOWESS is that the prediction and confidence intervals can not be calculated as easily as for the parametric regressions, they need to be assessed with re-sampling methods such as bootstrapping.



(a) $l = 0.3$ (b) $l = 0.8$

Figure 2.6: LOWESS fits on the data sample in Fig. 2.4 for two different span parameters l.

2.4 Gaussian Process (GP) Regression

The application of Gaussian Processes as a tool for nonlinear regression is the main subject of this work. Hence, in this Section a formal definition and an explanation how it is possible to obtain predictions with Gaussian Processes are given. Also it is demonstrated how the actual learning process with Gaussian Processes works. Finally the benefits and drawbacks of Gaussian Processes are outlined. The first studies of Gaussian Processes date back to the end of the 19th century as documented in Lauritzen (1981), related models developed in the 20th century include the work of Matheron (1963), better known as *Kriging* in Geostatistics, and O'Hagan (1978). More recently, the Gaussian Process framework was substantially extended by Rasmussen (1996); Neal (1997); MacKay (1998). This theoretical summary will follow mainly the descriptions in Bishop (2006) and Rasmussen & Williams (2006).

2.4.1 Definition

Attempting a verbal definition of a Gaussian Process (GP), there are several possibilities. One of the most intuitive ones is given by Bishop (2006), defining a Gaussian Process as a *distribution over functions*. One can also say that a Gaussian Process is a generalization of a multivariate Gaussian distribution to infinitely many variables (Rasmussen & Williams, 2006). For a better understanding of the latter definition, it is helpful to imagine a function as an infinitely long vector.
In a more formal way, a GP is defined as following (Rasmussen & Williams, 2006):
A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution.
Putting the formal definition of a GP into mathematical equations a mean function $m(x)$ and a covariance function $k(x, x')$ need to be defined, which together completely specify a GP:

$$f(x) \sim GP\big(m(x), k(x, x')\big), \tag{2.24}$$

with

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)], \\ k(x, x') &= \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]. \end{aligned} \tag{2.25}$$

Often the mean function $m(x)$ is assumed to be zero, for notational simplicity, although it can also be assigned another value. The mean function, together with the covariance function, expresses the prior beliefs of the distribution over functions and fixes the properties of the actual functions used for inference. Random samples drawn from a GP prior are shown in Fig. 2.7, whereas in formal terms a GP prior specifying e.g., the random variables $(y_1 \cdots y_{100})$ is given by

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{100} \end{bmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(x_1) \\ \vdots \\ m(x_{100}) \end{pmatrix}, \mathbf{K} \right), \tag{2.26}$$

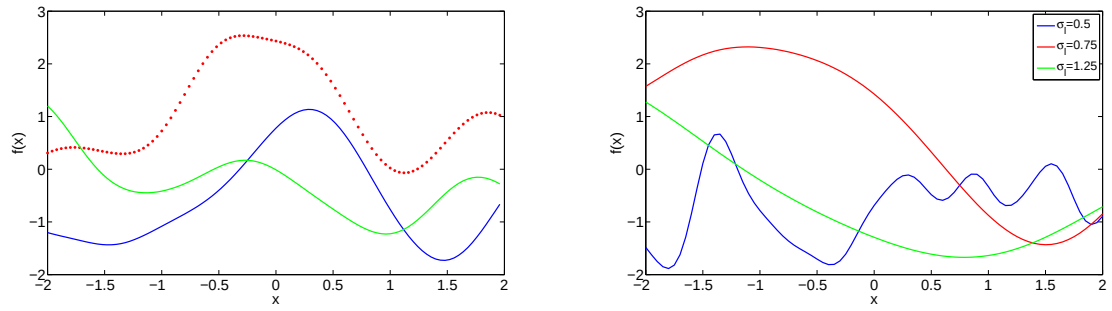
with

$$\mathbf{K} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_{100}) \\ \vdots & \ddots & \vdots \\ k(x_{100}, x_1) & \cdots & k(x_{100}, x_{100}) \end{pmatrix}. \tag{2.27}$$

The values in the covariance matrix K can be calculated by applying a covariance function of choice. The covariance function specifies the covariance between pairs of random variables. A very common example for a covariance function is the squared exponential covariance function, which is in fact a Gaussian kernel:

$$\mathrm{cov}(y) = k(x, x') = \sigma_f^2 \exp(-\|x - x'\|^2 / 2\sigma_l^2). \tag{2.28}$$

Note that the covariance between e.g. the ouputs y_1 and y_2 is a function of the inputs x_1 and x_2 exclusively. For this particular covariance function, only the signal variance parameter σ_f and the length-scale parameter σ_l, the latter defines the variance of the function along the x-axis, also have an influence on the final covariance. Another possible definition of the characteristic length-scale is "the distance one has to move in the input space before the function value can change significantly" (Rasmussen & Williams, 2006), thus characterizing the smoothness of the prior functions and the properties of the covariance function. Its influence is depicted in Fig. 2.7(b) where samples from the GP prior above are drawn, but with a varying length-scale parameter.



(a) same length-scale (b) different length-scale

Figure 2.7: Random samples drawn from a Gaussian Process prior with mean zero and a squared exponential covariance function. Panel (a) shows three sample functions with the same length-scale parameter σ_l, whereas the red dots show 100 points actually generated from eq. 2.26 and the two other functions are lines of joint points. In panel (b) the three sample functions show how different length-scales influence the smoothness of the functions.

2.4.2 Prediction with GPs

How can this prior distribution over functions be used to predict function values of new unseen data?

Given some noise-free observations $\{(x_i, y_i)|i = 1 \cdots n\}$ with f$(x) = y$ and f following a Gaussian distribution. Let x be a set of n training points of which the corresponding set y is known, and x^* be a set of test points for which the corresponding values y^* are to be predicted. Then the joint distribution of the training points and the test points is

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{K}(x,x) & \mathbf{K}(x,x^*) \\ \mathbf{K}(x^*,x) & \mathbf{K}(x^*,x^*) \end{bmatrix}\right). \tag{2.29}$$

In order to get the posterior distribution over functions, one needs to restrict that joint prior distribution to contain only those functions that agree with the observed training data points. This can simply be done by conditioning the joint Gaussian prior distribution on the observations, yielding for any test value y_{n+1} in y^* a Gaussian distribution $y_{n+1}|x_{n+1}, x, y \sim \mathcal{N}(m(x_{n+1}), \sigma^2(x_{n+1}))$ with the following mean and covariance

$$m(x_{n+1}) = \mathbf{K}(x_{n+1}, x)\mathbf{K}(x,x)^{-1}y, \tag{2.30}$$

$$\sigma^2(x_{n+1}) = \mathbf{K}(x_{n+1}, x_{n+1}) - \mathbf{K}(x_{n+1}, x)\mathbf{K}(x,x)^{-1}\mathbf{K}(x, x_{n+1}). \tag{2.31}$$

Note that the covariance of x_{n+1} is independent from the observed data in y, this is a property of the Gaussian distribution. However, the targets in y are included in the calculation of the new mean, which is a linear combination of y. Sampling from the posterior distribution can be realized by evaluating the mean and covariance matrix (which can simply be calculated by eq. 2.31 when more than one value is to be predicted), and computing the Cholesky decomposition of the covariance matrix. A Cholesky decomposition $\mathbf{L}$ of a positive-definite matrix $\mathbf{A}$ is defined as the decomposition of matrix $\mathbf{A}$ into $\mathbf{L}\mathbf{L}^T$, the product of a lower triangular matrix $\mathbf{L}$ and its transpose. The samples can then be calculated by $y_{sample} = m + \mathbf{L}x_{sample}$ at randomly drawn values x_{sample}. Note that the samples in Fig. 2.7 were drawn in the same way.
The samples in Fig. 2.8(a) show that by the calculation of the conditional distribution the functions have been modified so that they fit the observed data, whereas those functions in the prior distribution that disagree with the observations are in a way excluded in the posterior distribution. The mean of the functions in the posterior distribution along the y-axis (as calculated by eq. 2.30) represents (at point x_{n+1}) the actual predicted value y_{n+1} whereas the uncertainties of the prediction are represented by the pointwise double standard deviation for each input value (shaded areas in Fig. 2.8(b)). The uncertainty of the predictions can also be derived from the variance of the sampled functions just by visual inspection as one would keep on drawing samples from the posterior distribution as in Fig. 2.8(a). Notice that areas for which there are significant differences in the samples (such as for $x = [0.25, 0.5]$) have bigger uncertainties than areas where the samples are rather similar (such as for $x = [1, 1.25]$). In general, the predicted uncertainties get larger (magnitude dependent on the length-scale) in areas that are distant from any training points, which is in agreement with an intuitive understanding of data uncertainty. In order to approach more realistic scenarios one needs to consider data with noise, where it is assumed that the data y is created by some process $f(x)$ but is additionally influenced by some noise ϵ, as is generally the case when dealing with measured data. Usually additive noise is modeled as being independent and identically distributed Gaussian noise with variance σ_n^2, giving a model for the actual observations $z = y + \epsilon$. Therefore the covariance function of the prior including noise can be written as

$$\text{cov(z)} = \mathbf{K}(x, x) + \sigma_n^2\mathbf{I}. \tag{2.32}$$

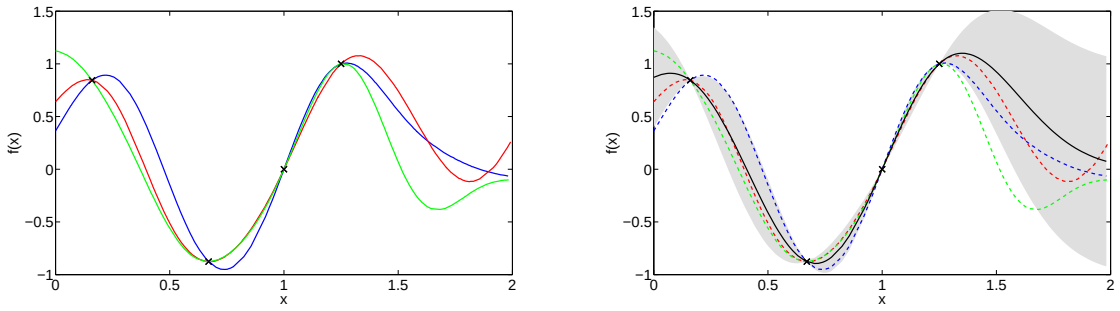


Figure 2.8: (a) Random samples drawn from a joint posterior distribution conditioned on four data points. (b) The same random samples as in a) (dashed lines) with the black line showing the mean of the distribution and the shaded areas showing the uncertainties by the two times standard deviation (corresponding to the 95% confidence region).

From eq. 2.29 one can thus infer the joint distribution of the observed data z and the function values y_* which are to be predicted as

$$\begin{bmatrix} z \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{K}(x,x) + \sigma_n^2\mathbf{I} & \mathbf{K}(x,x_*) \\ \mathbf{K}(x^*,x) & \mathbf{K}(x^*,x^*) \end{bmatrix}\right), \tag{2.33}$$

here also accounting for noise in the measurements, which means in practice that the data must not necessarily be fit exactly by the predicted functions anymore.
Applying this joint distribution to the predicted mean and covariance in the eq. 2.30 and eq. 2.31 one can define the new predictive equations as

$$m(x_{n+1}) = \mathbf{K}(x_{n+1},x)[\mathbf{K}(x,x) + \sigma_n^2\mathbf{I}]^{-1}y, \tag{2.34}$$
$$\sigma^2(x_{n+1}) = \mathbf{K}(x_{n+1},x_{n+1}) - \mathbf{K}(x_{n+1},x)[\mathbf{K}(x,x) + \sigma_n^2\mathbf{I}]^{-1}\mathbf{K}(x,x_{n+1}). \tag{2.35}$$

Note that the σ^2 here refers to the variance of the predicted function, thus it can be used to estimate confidence intervals. Based on σ^2 the calculation of the prediction intervals can be performed, by adding the estimated noise variance hyperparameter σ_n^2. Given the example case that the latent function is $f(x) = \sin(3\pi x)$, but the observed data contains only noisy versions thereof, which is $z = f(x) + \epsilon$. The noise follows a Gaussian distribution again. By variation of the noise level parameter σ_n^2 (which is added to the covariance function) one can show its influence on the resulting GPs when attempting to predict the latent function $f(x)$ (see Fig. 2.9). Whilst in Fig. 2.9(a), depicting a GP with

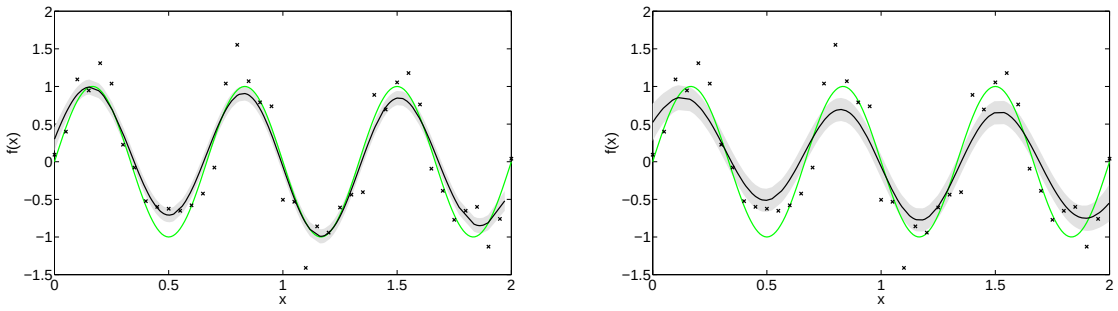


Figure 2.9: Influence of a varying noise level parameter on the resulting GP, keeping the length-scale parameter fixed. The green line depicts the latent function that is to be learned from noisy observations (x signs). (a) shows the resulting GP with $\sigma_n^2 = 0.05$ and (b) with $\sigma_n^2 = 0.15$, the black line again showing the mean function and the shaded areas the uncertainties.

a lower noise level parameter σ_n^2, the predicted function is closer to the observations, in Fig. 2.9(b) the higher noise level results in a GP that is smoother (referring to a slower variation), but also showing larger error bars for the uncertainties because it explains the distance of the predicted function to the actual data points by the noise level. Whereas it might seem reasonable by visual inspection that the function inferred in Fig. 2.9(a) explains the data more sufficiently than the function predicted in Fig. 2.9(b), this decision becomes more complicated for more complex or multidimensional data. Keeping in mind that the two above examples had a fixed length-scale parameter, but that this parameter can also be varied, it becomes clear that it needs powerful methods to estimate the parameter set that best explains the given data. Fortunately, one can use the principle of calculating the marginal likelihood $p(y|\theta)$ (eq. 2.36), which is the likelihood of the data given the (hyper-)parameters θ and refers to the marginalization over the latent noise-free function values. The marginal likelihood provides a measure to rank different models and parameter sets and therefore is an essential tool towards learning with GPs, as will be shown in the next Section.

2.4.3 Learning with GPs

The key idea of learning with Gaussian Processes is to define the covariance function properties in an appropriate manner to match the required application. This can mainly be done through optimizing the free parameters of the covariance function. The generally most important parameters of the squared exponential covariance function, i.e.

the length-scale σ_l, the signal variance σ_f and the noise level σ_n, have already been introduced before. Note that these parameters are referred to as *hyperparameters* by the GP pioneers Rasmussen and Williams, to stress that they are parameters of a non-parametric, probabilistic model. Optimizing these free hyperparameters can be done by e.g. maximizing the marginal likelihood (eq. 2.36) of the GP with a conjugate gradient-based optimization approach.
Moreover there are numerous covariance functions that can be chosen from, which are also part of the model selection problem. Some commonly-used covariance functions such as the Matérn covariance functions, exponential covariance functions or dot product covariance functions are introduced in this Section. However, the selection might be application specific and not complete.
The marginal likelihood forms a central element of learning with Gaussian Processes, with the term *marginal* refering to the marginalization of the latent function values. The probability of the data given the free hyperparameters is given by the log marginal likelihood (eq. 2.36).

$$\ln p(y|\theta) = -\frac{1}{2}y^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}y - \frac{1}{2}\ln|\mathbf{K} + \sigma_n^2\mathbf{I}| - \frac{n}{2}\ln 2\pi. \tag{2.36}$$

It might be misleading that the marginal likelihood is conditioned on θ, whereas it does not appear on the right hand side of the equation. Since θ is a vector containing all free hyperparameters which can be found in the covariance function, they contribute to the right hand side implicitly by influencing the final covariances in $\mathbf{K}$.
The marginal likelihood is formed by subtracting the complexity penalty $-\frac{1}{2}\ln|\mathbf{K}+\sigma_n^2\mathbf{I}|$ from the data fit term $-\frac{1}{2}y^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}y$ and a normalization constant $-\frac{n}{2}\ln 2\pi$. The complexity term accounts for less complex models depending on the covariance matrix only, whilst the data fit term includes the observed function values y and decreases with the length-scale, because the longer the length-scale the less flexible the model (Rasmussen & Williams, 2006) in terms of fitting the data.
In order to learn the hyperparameters that best fit the data, the marginal likelihood needs to be maximized by seeking the partial derivatives of the log marginal likelihood w.r.t. the hyperparameters (eq. 2.37).

$$\frac{\partial}{\partial\theta_j}\ln p(y|\theta) = -\frac{1}{2}y^T\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\theta_j}\mathbf{K}^{-1}y - \frac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\tfrac{\partial\mathbf{K}}{\partial\theta_j}). \tag{2.37}$$

The principle of maximizing the marginal likelihood has the advantage that it is analytically tractable, in contrast to other methods such as Bayesian principles where integrals over the parameter space are intractable. Of course the gradient based method still suffers from the possibility of multiple local optima. And in fact every local optimum corresponds to a different interpretation of the data, making it a crucial goal of learning with GPs not to get stuck in a local maximum. However, according to Rasmussen & Williams (2006), this should only occasionally be a problem if datasets are too small. With data sets of a sufficient size "'one often finds that one local optimum is orders of magnitude more probable than other local optima"'. Another possibility is, nevertheless, averaging together different explanations of the data.

2.4.4 Heteroscedastic GPs

When it can be assumed that the observations are influenced by input-dependent noise levels, Heteroscedastic (referring to a non-homogeneous noise variance) Gaussian Processes (HGPs) are the method of choice.
The concept of HGPs was first introduced by Goldberg *et al.* (1998). Assuming that the noise is a smooth function of the inputs, it can be modeled using a second independent Gaussian Process (the z-process), whilst the actual function values are assumed to be noise-free and modeled with a standard GP (the y-process). Making predictions with HGPs is not as easy with as with standard GPs, because the predictive posterior distribution now also accounts for the noise rates as independent latent variables. This, in turn makes the integral for the predictive posterior distribution difficult to handle analytically. Therefore Goldberg *et al.* (1998) proposed a Markov Chain Monte Carlo approach to sample from the distribution of the latent noise variables. In contrast, Kersting *et al.* (2007) used a most likely noise approach to approximate the posterior noise variance.
The principle used by Kersting *et al.* (2007) is similar to the one underlying the (hard) Expectation Maximization Algorithm (Dempster *et al.*, 1977): if the values of the noise levels are actually known, than learning the parameters is easy. The iterative algorithm for learning the hyperparameters of both the y-process and the z-process concurrently with the HGP approach can be summarized in 5 steps (Algorithm 2.1).

Algorithm 2.1 Most likely noise approach for a HGP

1: Estimate a homoscedastic GP G1 that maximizes the likelihood of the data.
2: The empirical noise levels are estimated for the training data, by minimizing the average distance between the predictive distribution given by G1 and the prototype value (derived by repeated sampling from the predictive distribution at a given data point)
3: A second GP G2 is estimated on the empirical noise levels.
4: G1 is combined with G2 to estimate GP G3, with which the noise levels of interest can be predicted.
5: If not converged, G1 is set to G3 and the steps 2-4 are repeated.

This approach is not guaranteed to improve the likelihood in every iteration, but, in case of oscillations, will most often stop improving at reasonable parameter estimates. Regarding noise estimates, it outperformed standard GPs in a variety of examples (Kersting *et al.*, 2007) and is competitive with other heteroscedastic regression approaches (e.g., Schölkopf *et al.* (2000)).
A challenging feature of the HGPs is that the parameter n_z, which determines the number of the latent noise variables, needs to be optimized manually. Good picks are usually in the range of $n_z = [3,20]$.

2.4.5 Benefits and drawbacks

Gaussian Processes provide a framework to deal computationally (in terms of tractability) with inference along infinite dimensional objects. They can be applied both for regression and classification evaluate confidence and prediction intervals simply by the posterior distribution. Moreover, it is trivial to extend the predictions as presented in 2.4.2 to multidimensional inputs, only the covariance function must be adapted so that it can evaluate several multidimensional x-values. However, since the matrix inversion of a $n \times n$ matrix in eq. 2.30 and 2.31 is only possible in $O(n^3)$, GPs can have problems with huge data sets. Things usually start to get difficult for $n > 10000$.
Another difficulty with Gaussian Processes is the selection of a suitable covariance function. There are endless possibilities in combining different covariance functions to composites, which then can model very complex behaviour. There is no automatic way for finding the "most appropriate" covariance function for a specific problem. Nevertheless, covariance functions offer the possibility to incorporate prior knowledge, if available, into the GP inference.

Under the assumption of Gaussian observation noise, the computations needed to make predictions with GPs are analytically tractable (Rasmussen & Nickisch, 2010). In cases where this not holds true, nonlinear optimizations to find the hyperparameters need to be applied and one is confronted with the same local minima issues as in other learning methods.

MacKay (1998) gives an interesting discussion towards the relationship of Gaussian Processes to other Machine Learning methods such as Artificial Neural Networks (ANNs), regarding the definition of Gaussian Processes as "ANNs with infinitely many hidden units". One might argue that Gaussian Processes are just smoothing devices, lacking the ability to reveal the hidden features of a problem in a mathematically enclosed form. Thus, e.g., feature discovery is a task that not should be adressed by Gaussian Processes. On the other hand, there are many real world data problems which can be solved by sensible smoothing methods, and so with Gaussian Processes (Williams & Rasmussen, 1996; Gao, 2004; Sun *et al.*, 2010; Macke *et al.*, 2011).

3 Results

This diploma thesis is mainly a methodological project, focused on the theory and applicability of Gaussian Processes. To investigate and point out their strengths and weaknesses, numerous experiments including different setups have formed the central piece of work. In this chapter, results on several simulated data sets (Sections 3.2 and 3.4) are visualized, quantified and discussed. Knowing the expected outcomes beforehand, the artificial experiments provide a good basis for a comparison to the performance on the according real world datasets, i.e. actually measured ecosystem data (Sections 3.3 and 3.5). The possibilities of simulating data are endless, this is why the selection of experiments discussed here represent rather a snapshot of potential simulation work.

3.1 Application Specific Methods

3.1.1 General

Throughout this chapter, the two different GP approaches (described in more detail in Section 2.4) will be distinguished as "GPMLs", if the experiments were performed using exclusively the `MATLAB` *gpml toolbox* (Rasmussen & Nickisch, 2010) and as "HGPs", if code for *most likely heteroscedastic GPs* developed by Kersting *et al.* (2007), which is also based on the *gpml toolbox*, was applied.
For both GP approaches the squared exponential covariance function, with the initial hyperparameters set to $\sigma_l = \sigma_f = \sigma_n = 1$, was used. Since the length-scale parameter σ_l is arguably the most essential hyperparameter amongst them, it was modified after optimization, such that its statistical modelling power could be investigated. Given the respective optimal GPs as $\mathrm{GPML_O}$ and $\mathrm{HGP_O}$, their modified counterparts are defined as $\mathrm{GPML_S}$ and $\mathrm{HGP_S}$ for the optimal σ_l reduced by $e^{1/2}$ ("shorter length-scale") and $\mathrm{GPML_L}$ and $\mathrm{HGP_L}$ for the optimal σ_l increased by $e^{1/2}$ ("longer length-scale").

In the following experiments the standard NLR method (Section 2.3.1), with a prescribed function given and learning the optimal parameters directly via nonlinear least squares, will be referred to as the *reference method*. The fit routine `nlinfit.m` from the `MATLAB` statistics toolbox was applied here. Knowing the form of the actual data generating function beforehand, NLR cannot be considered as a benchmark method per se. Therefore a LOWESS approach (Section 2.3.2) was used as the *benchmark method*, to compare the GPML and HGP results with another non-parametric nonlinear regression method. A weighted linear least squares with a tri-cube weight function and a first degree polynomial model for accomplishing the smoothing with local regressions is applied, as implemented in the functions `smooth.m` and `fit.m` included in the `MATLAB` curve fitting toolbox. Uncertainty esimates for both the NLR and the LOWESS fits are provided via bootstrapping, with the number of bootstrap samples set to 999.

3.1.2 Light Response Data

The relationship of the carbon exchange between biosphere and atmosphere to its climatic controls has long attracted researchers from different backgrounds (e.g., Smith (1938), Baldocchi (1997) or Bonan (2002)). In ecophysiology, the carbon exchange is usually measured as the Net Ecosystem Production (NEP), which equals the gross primary production (GPP) minus the ecosystem respiration (R_{eco}).
Recently, in an extensive Artificial Neural Network (ANN) study (Moffat, Accepted), NEP was modeled as the response to 25 radiative, meteorological and theoretical driver candidates. It could be confirmed that $PPFD$ (Photosynthetic Photon Flux Density) is the most relevant of them for the daytime during the growing season.
The Smith Sigmoid function f_{SS} is known to be a suitable representation for the light response (Moffat, Accepted):

$$f_{SS}(PPFD) = \frac{\alpha \cdot GPP_{opt} \cdot PPFD}{\sqrt{GPP_{opt}^2 + (\alpha \cdot PPFD)^2}} - ER_d. \tag{3.1}$$

This light response curve is characterized by three parameters (α, NEP_{sat}^* and ER_d). First the quantum yield parameter

$$\alpha = \frac{df_{SS}(PPFD)}{dPPFD} \text{ at small } PPFD, \tag{3.2}$$

which represents a linear increase in the light response at low $PPFD$, before the curve gradually curves to a maximum photosynthesis rate. α can be estimated by the initial slope (i.e. the first derivative) of the curve.

$$\frac{df_{SS}(PPFD)}{dPPFD} = \frac{\alpha \cdot GPP_{opt}^3}{(\alpha^2 PPFD^2 + GPP_{opt}^2)^{3/2}}. \tag{3.3}$$

These values can then be compared to the numerical pointwise derivatives of the predicted GP functions, approximated by

$$f'_{SS}(PPFD) \approx \frac{f_{SS}(PPFD + h) - f_{SS}(PPFD)}{h}, \tag{3.4}$$

with h chosen to be a small number close to zero, since the derivative of a function f at x_0 is given as (e.g., Burden & Faires (2000))

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}. \tag{3.5}$$

Next, the respiration parameter

$$ER_d = -NEP(0), \tag{3.6}$$

which is the intercept at zero $PPFD$ when there is no photosynthetical uptake. Third, the optimum gross primary production parameter

$$GPP_{opt} = NEP_{sat} + ER_d, \tag{3.7}$$

with

$$NEP_{sat} = NEP(PPFD_{sat}) = const, \tag{3.8}$$

is defined as the light saturated photosynthesis rate. It is unknown at what light intensity an ecosystem saturates, hence there can be different characterizations of NEP_{sat}. For instance the Smith sigmoid function saturates at $PPFD$ towards infinity, hence it models the saturation of photosynthesis asymptotically. For reasons of simplification NEP_{sat} is set to NEP at the maximum measured light intensity here:

$$NEP^*_{sat} = NEP \text{ at maximum } PPFD. \tag{3.9}$$

This assumption also provides a good basis for comparing non-parametric with parametric models, since NEP^*_{sat} can be inferred from the (fitted) curve.

3.1.3 Algorithm for Rb estimation

The ecosystem respiration R_{eco} can be modeled by the Lloyd-Taylor model

$$R_{eco} = Rb \cdot \exp\Big(E_0 \cdot \Big(\frac{1}{T_{ref} - T_0} - \frac{1}{T_a - T_0}\Big)\Big). \tag{3.10}$$

For the estimation of the Rb parameter (base respiration at reference temperature T_{ref}), the following simple 5-step algorithm was developed:

Algorithm 3.1 R_{eco} via Rb estimation

1: Train a HGP on (incomplete) annual flux measurements with drivers $[PPFD, T_a, time_d]$.
2: Predict annual NEP using the annual meteorological time series from the golden file.
3: Estimate the Rb parameter at $PPFD = 0, T_a = 15°$ along annual $time_d$.
4: Set the parameters in the Lloyd-Taylor model to:
$T_{ref} = 15°\text{C}$
$T_0 = -46.02°\text{C}$
Incorporate the E_0 estimates of Reichstein *et al.* (2005) into the model.
Calculate annual R_{eco} for the measured whole year T_a time series from the golden file.
5: Calculate $GPP = R_{eco} + NEP$.

3.2 Simulated Light Response Data (SLR)

3.2.1 Data

To simulate data that resembles the light response of a beech forest and has known properties, the Smith Sigmoid function f_{SS} (eq. 3.1) was used as a generating function. Since there is always some systematic and some random noise in measured data (Section 2.1.3), the values given by eq. 3.1 (parameterized following Moffat (Accepted), see text

box in plot) would not be observed in real world ecosystems. To simulate noise with a realistic magnitude σ, disturbance terms derived from ANN model residuals binned by the NEP flux magnitude in steps of 5 μmol CO_2 m^{-2} s^{-1} (Moffat, Accepted) with the standard deviation

$$SDev_{ANN} = 2.5(\pm 0.3) + 0.11(\pm 0.02) \cdot NEP. \tag{3.11}$$

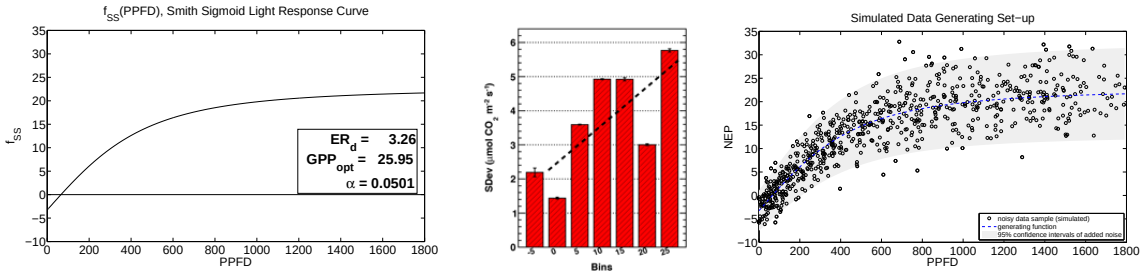
This was added to NEP as

$$\sigma = 2.5(\pm 0.3) + 0.11(\pm 0.02) \cdot f_{SS}, \tag{3.12}$$

resulting in

$$NEP = f_{SS}(PPFD) + \sigma. \tag{3.13}$$

First, to keep things simple, Gaussian white noise was simulated, which can be done by repeatedly drawing normally distributed pseudorandom numbers. The sampled data is not equally distributed but more sparse with increasing flux magnitude. The non-equal distribution of the total of 721 data points was simulated by using the frequencies of the data points derived from a measured daytime data set from three successive summers (sample used in Section 3.3), with frequencies taken over an input range of 200 $PPFD$.



(a) Smith Sigmoid (b) noise (c) simulated data

Figure 3.1: Setup of simulated data. (a) Smith sigmoid light response function, parameterized with ANN derived parameters as shown in the box, (b) linear regression on the standard deviation of the ANN model residuals binned over a NEP magnitude of 5 μmol CO_2 m^{-2} s^{-1}, used as a noise model (c) resulting simulated data sample.

3.2.2 Performance

All GP and HGP runs as well as NLR and LOWESS had an R^2 performance of 79%(±1%), the corresponding value of the generating function was 79% (see Table 3.1). Some of the algorithms such as GPML_S and HGP_O even did slightly better than the generating function, which might be due to fitted noise levels. The maximum deviation of ±0.03 in the $RMSE$ between all methods suggests that all of them have a very similar performance in terms of a statistic quantity and provided a good fit of the data.

method	$\mathbf{R^2}$	**RMSE**	**bias**	α	$\mathbf{ER_d}$	$\mathbf{NEP^*_{sat}}$
data	**0.79**	**3.99**	**-0.12524**	**0.0501**	**3.26**	**21.67**
NLR	0.79	3.98	-0.00000	0.0491 *0.0010* (± 0.0044) *2.1* (± 9.0)%	3.20 *0.06*(± 0.63) *2.0* (± 19.0)%	22.14 *0.47*(± 0.70) *2.2* (± 3.0)%
LOWESS	0.80	3.96	-0.10006	0.0444 *0.0057* *11.4* %	2.92 *0.34*(± 0.66) *10.4* (± 20.0)%	22.35 *0.69* (± 1.53) *3.2* (± 7.0)%
$\mathbf{GPML_S}$	0.80	3.98	-0.00253	0.0481 *0.0020* *4.1* %	3.05 *0.21* (± 1.15) *6.6* (± 35.0)%	21.97 *0.30* (± 2.61) *1.4* (± 12.0)%
$\mathbf{GPML_O}$	0.79	3.98	-0.00026	0.0462 *0.0039* *7.8* %	2.81 *0.45* (± 0.94) *13.8* (± 29.0)%	22.58 *0.91* (± 2.11) *4.2* (± 10.0)%
$\mathbf{GPML_L}$	0.79	4.00	0.00156	0.0429 *0.0072* *14.3* %	2.23 *1.03* (± 0.79) *31.7* (± 24.0)%	22.57 *0.90* (± 1.65) *4.2* (± 8.0)%
$\mathbf{HGP_S}$	0.80	3.98	-0.00006	0.0461 *0.0040* *7.9* %	2.91 *0.35* (± 1.21) *10.9* (± 37.0)%	21.99 *0.33* (± 1.02) *1.5* (± 5.0)%
$\mathbf{HGP_O}$	0.80	3.98	0.00396	0.0471 *0.0030* *6.0* %	2.77 *0.49* (± 0.99) *15.0* (± 30.0)%	22.00 *0.34* (± 1.52) *1.5* (± 7.0)%
$\mathbf{HGP_L}$	0.79	4.00	0.00992	0.0458 *0.0043* *8.6* %	2.31 *0.95* (± 0.83) *29.3* (± 25.0)%	22.20 *0.53* (± 2.26) *2.5* (± 10.0)%

Table 3.1: Fit performance and physiological parameter estimates of all methods. Absolute deviations and deviations as a percentage are listed for each method in the rows below. Uncertainties are given in brackets by the double standard deviation referring to 95% confidence intervals. For NLR and LOWESS, the standard deviation was estimated by bootstrapping with a sample size of 999. Note, no GP uncertainties for α can be listed due to methodological limitations.

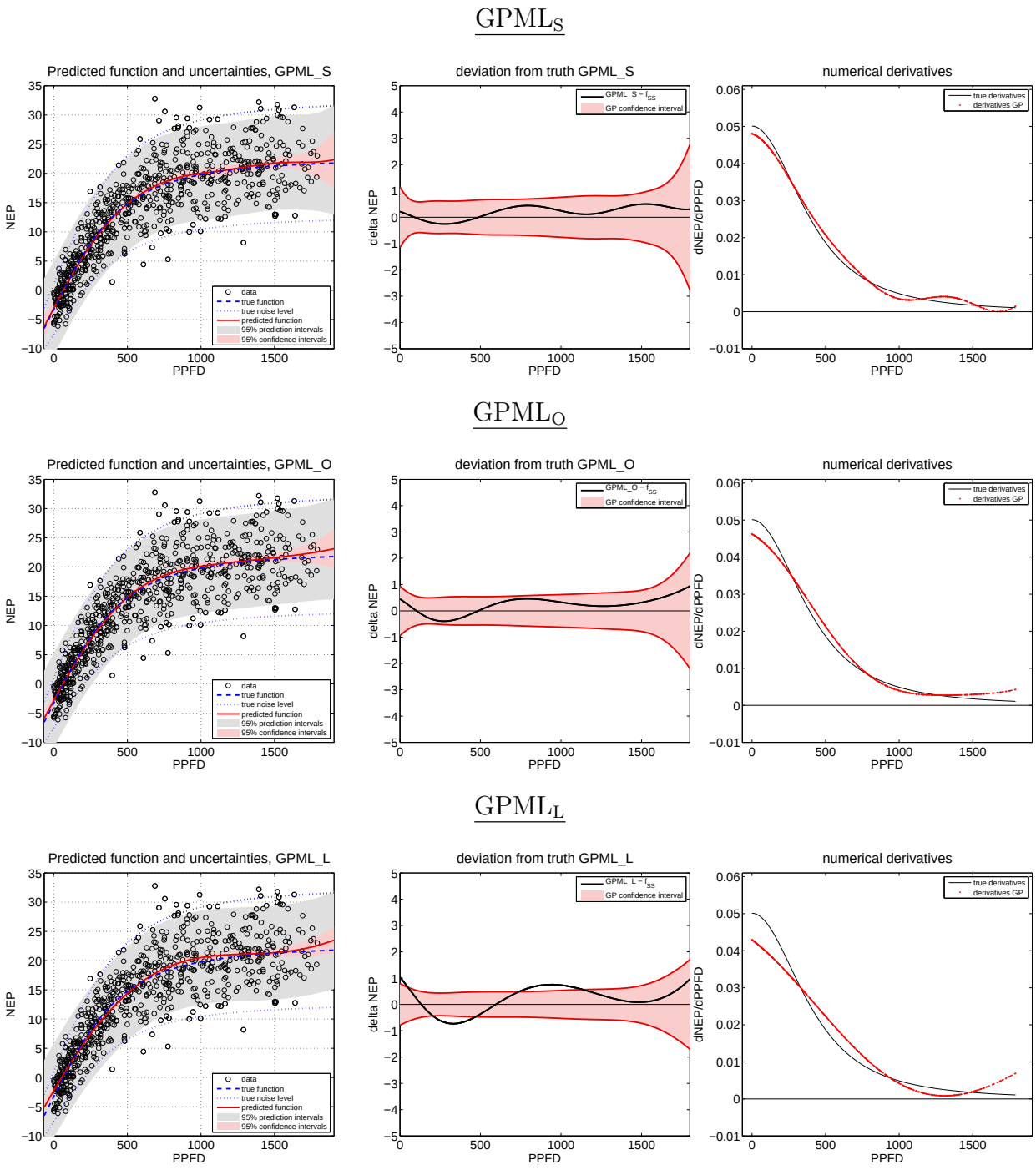


Figure 3.2: GPML prediction on the data sample shown in Fig. 3.1 for three different length-scale parameters.
Left: Fit performance, confidence and prediction intervals. **Center:** deviation of the predicted function towards the true function and the predicted confidence intervals. **Right:** First derivative of the Smith sigmoid function and the numerical pointwise derivatives of the respective GPML.
For an explanation of lines and shaded areas refer to the legends inside the axes.

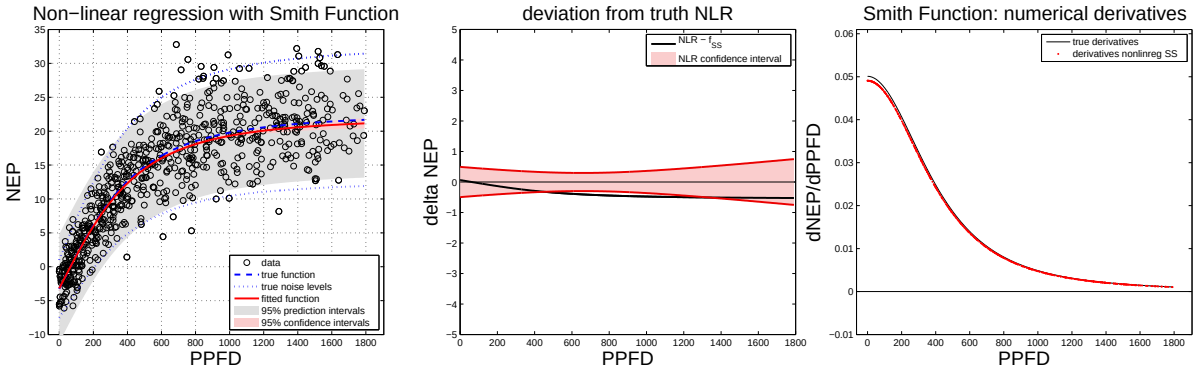


Figure 3.3: NLR fit on the data sample in Fig. 3.1. For the full caption of this plot refer to Fig. 3.2.

There are some fundamental differences in the actual fitted curves that can only be revealed when taking a closer look at the predicted functions and its derivatives. These will also be of importance when evaluating the GP's ability to estimate the physiological parameters and their uncertainties.

In this analysis, the parameters of the data generating set-up (cf. row "data" in Table 3.1) are referred to as the "truth". Arguably, one might also consider the parameters estimated by the NLR (Fig. 3.3) as the true parameters, since there is in fact no better predictor than the NLR which makes use of the generating function as prior knowledge. In other words, the NLR parameter set is the one that best fits the data sample drawn. However, the interest in this study is more in the underlying variance than in the sample-induced bias, which results in this understanding of the true parameters.

Analyzing the predicted functions, the magnitude of deviations from the true generating function is different for some of the fitted curves (Fig. 3.2). Whereas GPML_O and GPML_S show small deviations ($[-0.3, 0.8]\,NEP$ and $[-0.2, 0.6]\,NEP$, respectively), the deviations of GPML_L ($[-0.8, 1.1]\,NEP$) are higher and are even outside the confidence intervals. This can be explained by the longer length-scale parameter of GPML_L which results in less freedom for the predicted function and in combination with a smaller predicted confidence interval this causes the largest deviations. In contrast, the shorter length-scale parameter (GPML_S) gives a higher predictive freedom and results in a function that shows the most frequent turning points. In general, all of the predicted GPML functions "wobble" to some extent around the true underlying function. The LOWESS fit (Fig. 3.4) is generally comparable to GPML_O and GPML_S, showing only small deviations around $[-0.2, 0.6]\,NEP$ and a very similar curve. However, it becomes apparent that the fit is not as smooth as the GPML fits, as caused by the piecewise

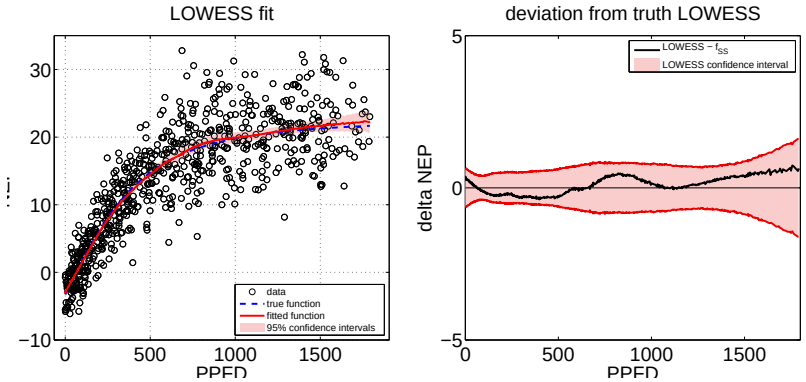


Figure 3.4: LOWESS fit on the data sample in Fig. 3.1. For the full caption of this plot refer to Fig. 3.2. The derivative cannot be calculated from the LOWESS fit.

local linear regressions.

Note that all the three GPML approaches have a positive deviation from the truth towards zero $PPFD$ and towards maximum $PPFD$. At zero $PPFD$, GPML_S shows small deviations (0.1 NEP), GPML_O shows medium deviations (0.6 NEP) and GPML_L has the largest deviations (1 NEP). Also at high $PPFD$, GPML_S proves to have the smallest deviations (0.2 NEP) among all GPMLs. Therefore, it is suggested that NEP is overestimated in these areas, a valuable information for parameter estimation of ER_d and NEP^*_{sat}. Looking at the derivatives of the predicted GPs and comparing them with derivative of the Smith sigmoid function (eq. 3.3) the results are likewise. In contrast to the derivative of the generating function, which shows a plateau at low $PPFD$, the pointwise numerical GPML derivatives are not constant in that region. Moreover, they also do not level off towards maximum $PPFD$. This indicates that the GPML method shows weaknesses in capturing certain local features of the function. GPML_S has the derivative that intersects the true derivative the most often, with five interSections it is "more wobbly" than its two counterparts that only cross the true derivative three times. Again, the more frequent changes in the slope of the predicted function can be explained by the different length-scales. Accordingly, the first derivative demonstrates the impact of the length-scale in a coherent way, i.e. with a longer length-scale parameter also the curvature of the first derivative gets smoother. In contrast, a shorter length-scale corresponds to a more unstable first derivative, showing that there is more variance along the x-axis of the predicted function.

As would be expected from a methodological point of view, the fit performance and the fitted functions of the according HGP fits, depicted in Fig. 3.5, resemble the ones of the GPML method, showing comparable length-scale effects and suffering similar

drawbacks concerning local features of a function. Hence, the HGP results are not discussed separately in this Section.

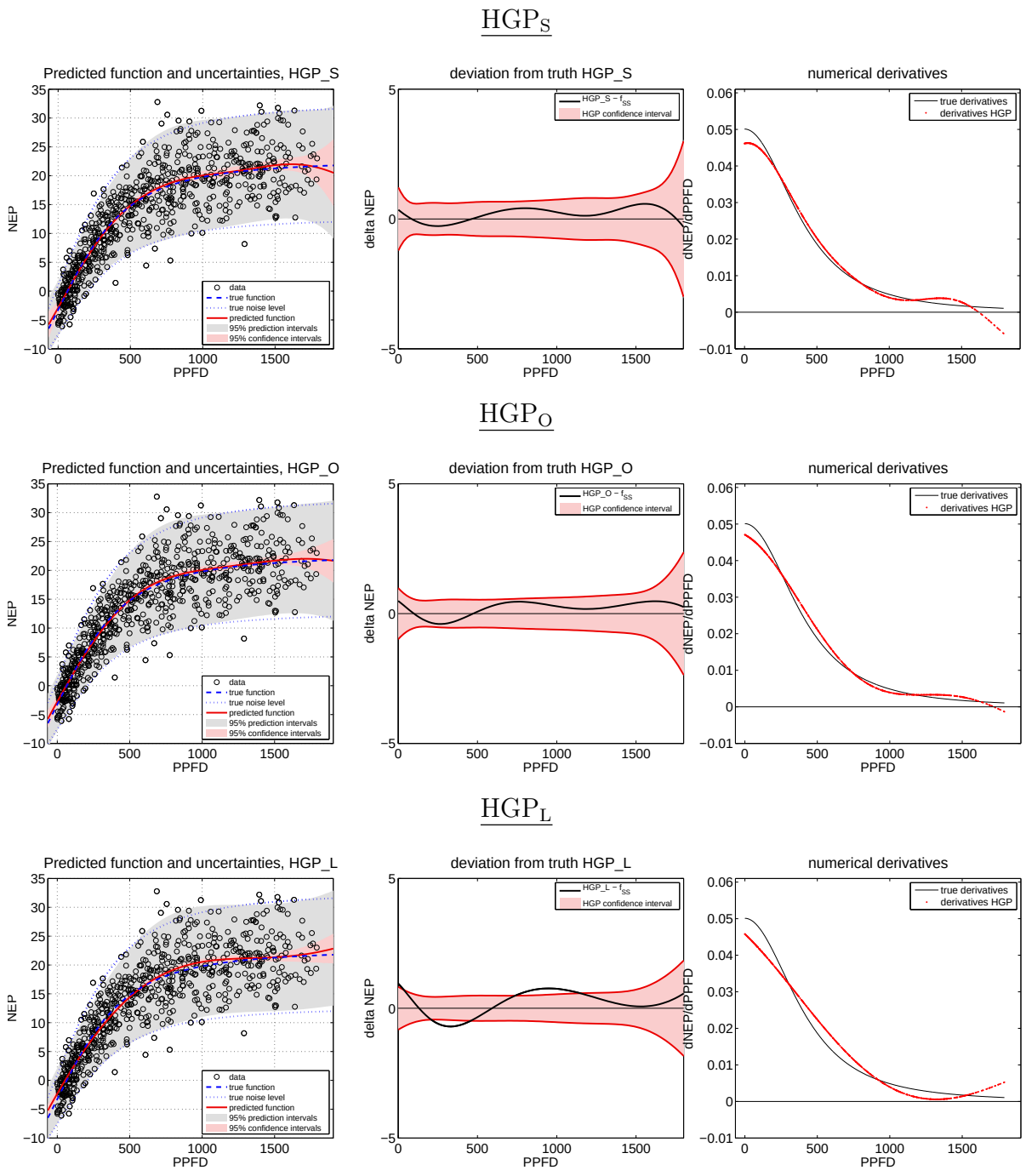


Figure 3.5: HGP prediction on the data sample shown in Fig. 3.1 for three different length-scale parameters. For the full caption of this plot refer to Fig. 3.2.

3.2.3 Physiological Parameters

Like decribed in Section 3.2.1, the three parameters of the Smith sigmoid function can be inferred from the curve of the (predicted) function. This is necessary in particular for the Gaussian Process methods, since they are a non-parametric model doing predictions based on the posterior distribution rather than on a mathematical closed term expression. In this Section, the parameters predicted by the GPs will be compared with the true parameters of the generating function based on Table 3.1. Note that in Table 3.1 the parameters for the NLR are derived accordingly (for better comparability) and are not the directly estimated regression parameters.

The initial quantum yield parameter α is underestimated within the range of 15% by all methods. Whereas GPML_L estimates the parameter with the highest deviation of 14.3%, all other GP approaches are between $4 - 9\%$ error and in a similar range as LOWESS (Fig. 3.3, 11.4% error). As can be expected, the NLR (2.1% deviation) as the reference method outperforms all other methods.

The respiration at daytime, ER_d, is also underestimated by all methods, which is not surprising since α and ER_d are both parameters estimated at zero $PPFD$ and therefore related. The expectation that was confirmed for the two parameters. Again, the NLR (2.0% deviation) outperforms the other methods.

The values in Table 3.1 suggest that the GPs do a better job for the estimation of NEP^*_{sat}, the approximated net ecosystem exchange at light saturation. All methods overestimate the parameter with at most 4.5% error towards the true parameter. Both GPML_S and HGP_S have the smallest error with 1.4% and are even slightly better than the reference method (2.2%) and the benchmark method (3.2%). However, this result needs to be treated carefully, since the predicted function does not saturate at high $PPFD$ (Fig. 3.2 and Fig. 3.5) as indicated by a first derivative that does not level off to zero. In the regions beyond available data, there is a very different behaviour in the predicted function for e.g. $\text{HGP}_\text{O}$and HGP_S. Thus, the good parameter estimates are to a high extent dependent on the (most sparse along the x-axis) data distribution in that area and sensitive to small changes in the length-scale parameter of the GP. In fact, a larger number of data points in that region would probably give a fit which is more robust.

Here, the values of the estimated parameters prove to be a good example of why it is better to compare NEP^*_{sat} instead of the parameter GPP_{opt}, which parameterizes the generating Smith sigmoid function. When underestimating ER_d and concurrently overestimating NEP^*_{sat}, the resulting compensation of errors might give misleading results.

Notably, the result of $GPML_S$ and HGP_S indicate that there could be a better fit "somewhere between e.g. $GPML_O$ and $GPML_S$". However, if the selected model class is chosen too big, this might lead to overfitting. On the other hand, the modified GPs with a longer length-scale give usually the worst results.
In general, both the GPML method and HGP method did a reasonable job in estimating the introduced parameters and overall there are only marginal differences towards the LOWESS approach. Consequently, the predicted confidence intervals of the estimates will be of interest, which is the subject of the next Section.

3.2.4 Confidence Intervals

The confidence intervals of the GP prediction (represented by the light red shaded areas) correspond to confidence bounds for the predicted mean function and are calculated by substracting the noise variance hyperparameter σ_n needs from the predictive variance (eq. 2.31) at every point of interest.
To evaluate the reliability of the confidence intervals, they were compared with the actual deviations from the true function, as shown in the right plots in Fig. 3.2. By this,

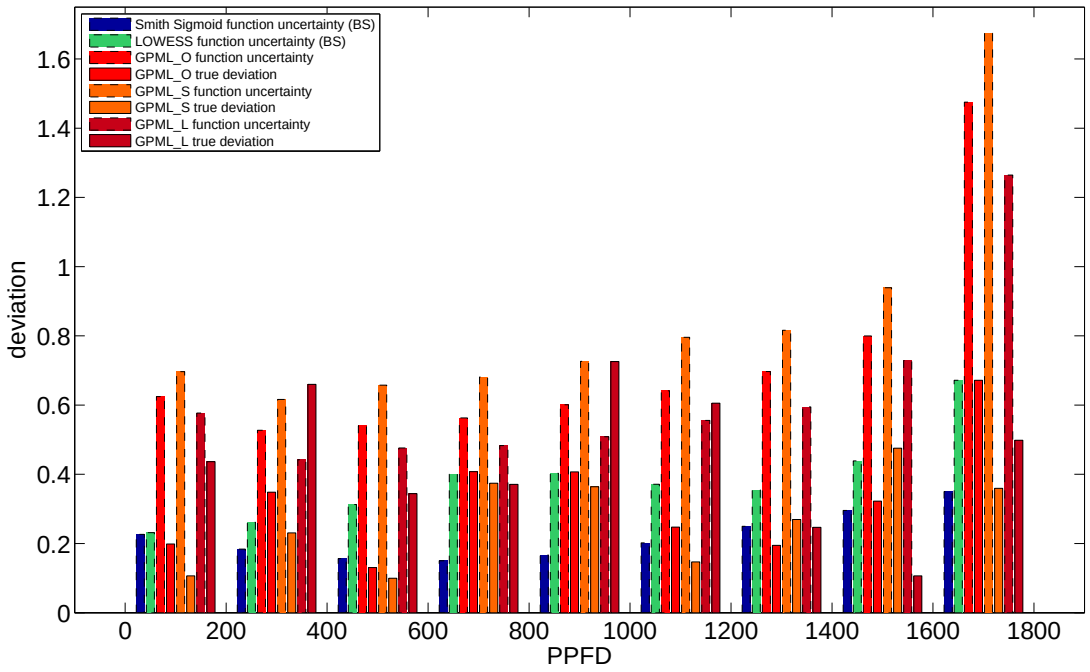


Figure 3.6: Binned GPML confidence intervals compared with the mean binned deviation of the predicted function to the true function.

areas where the predicted function exceeds the confidence intervals can be detected. In the experiment here, such areas occur only for the $\mathrm{GPML_L}$ prediction and not for $\mathrm{GPML_O}$ and $\mathrm{GPML_S}$. However, for the latter two, the deviations are sometimes close to the boundaries of the 95% confidence intervals, suggesting that the confidence intervals might not be reliable.
Another way of assessing the reliability of the predicted confidence intervals is to compare them with the mean deviation of the binned differences between the predicted and the generating function, as shown in Fig. 3.6. For $\mathrm{GPML_O}$ and $\mathrm{GPML_S}$, the mean deviations are always about half or less than the value of the predicted 95% confidence interval. Only in the bins $[600, 800]$ $PPFD$ and $[800, 1000]$ $PPFD$ the mean deviations exceed half the value of the predicted uncertainties. From this point of view, the confidence intervals do not necessarily need to be treated as unreliable. Nevertheless, it became apparent that they are very sensitive to the optimized length-scale parameter.
The confidence intervals for the HGP method are not part of the original HGP code. They were implemented to be simply the confidence intervals of the first, homoscedastic GP. This results in confidence intervals very similar to their GPML equivalents and for brevity they are not discussed separately here.

3.2.5 Prediction Intervals

The GPML and the HGP method can provide prediction intervals by calculating the variance of the predictive posterior distribution. For normal distributed data, the double standard deviation refers to 95% confidence intervals for predicted data, which will be called the "predictive uncertainties" and corresponds to the boundaries in which predicted data is to be expected. The prediction intervals for NLR was calculated as described in Section 2.3, whereas the LOWESS method cannot provide these intervals. Predictive uncertainties are depicted as gray shaded areas in all plots.
In the following the term "true residual" will be used to describe the difference between a sampled datapoint and the generating function value at that point (cf. the definition of a residual in eq. 2.4). Residuals are of interest here to assess the model error and for comparison with the GP prediction intervals.

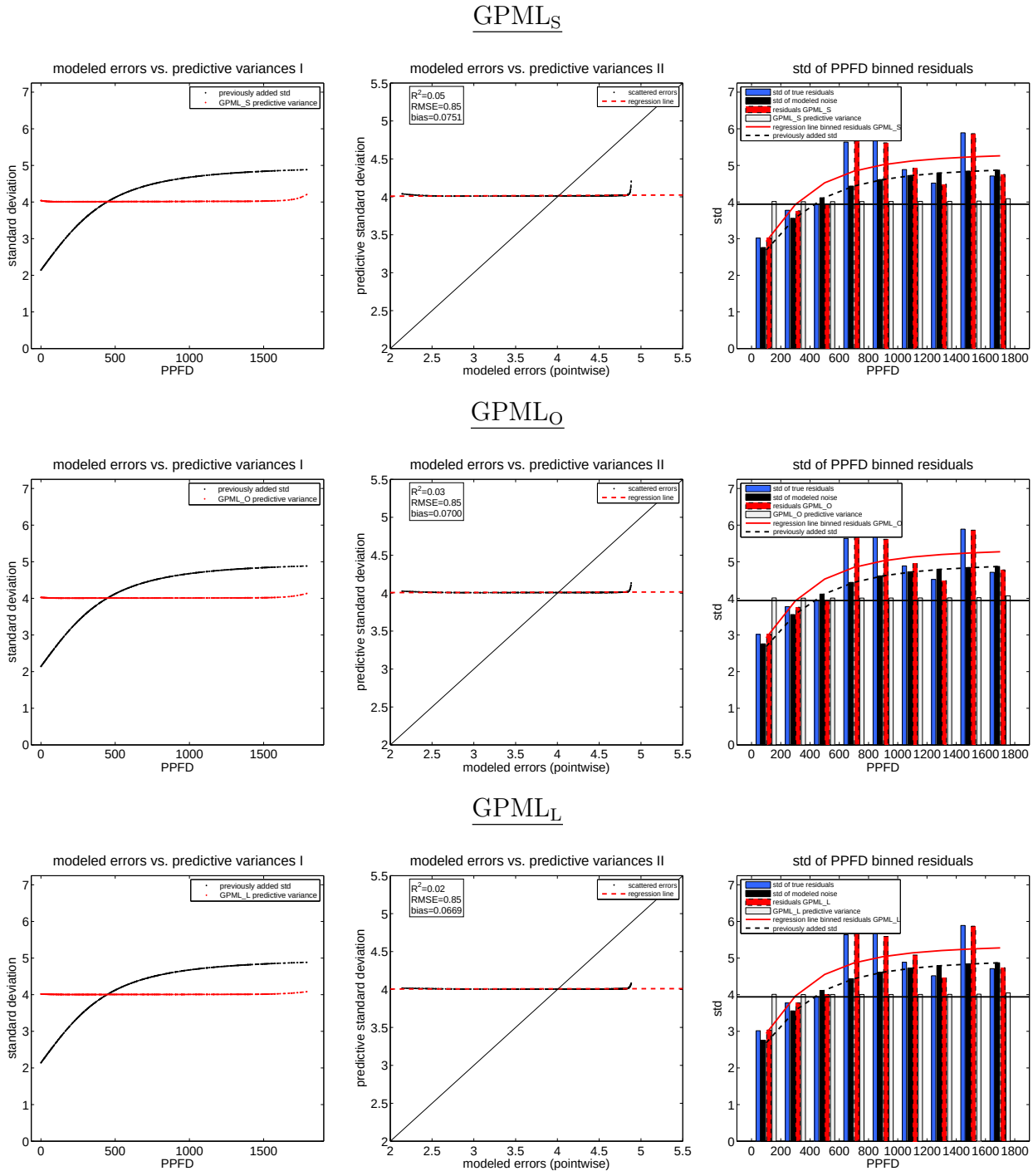


Figure 3.7: GPML predictive uncertainty analysis on the data sample shown in Fig. 3.1. **Left:** The modeled errors (i.e. the noise standard deviation σ) compared to the predictive variances σ_{GPML} against $PPFD$. **Center:** Scatterplot of the modeled noise vs. the predicted noise. **Right:** Binned model residuals compared to binned predictive variances and binned true residuals.
For an explanation of lines and bars refer to the legends inside the plots.

Unlike in the previous Sections, it is clearly distinguished between the GPML results from the HGP results since GPML and HGP behave differently.
The predicted standard deviation of all three GPMLs is nearly constant along the x-axis at 4 μmol CO_2 $m^{-2}s^{-1}$ (Fig. 3.7, left), which is very close to the mean of the previously added noise at 3.94 μmol CO_2 $m^{-2}s^{-1}$. The varying noise level could not be captured at all, with an $R^2 \leq 0.05$ in the relationship between modeled errors and predicted variances for the three GPML approaches, but also for the NLR. The discrepancies are also depicted in the performance plots (Figures 3.3 and 3.2, left) as the differences between the blue dotted lines and the gray shaded areas. However, the mean noise magnitude could be reproduced, what becomes apparent when comparing the gray bars to the straight black line depicted in the right plot(s) in Fig. 3.7. This is an interesting result and further investigation should be done, e.g. it needs to be shown if the result can be generalized along different noise magnitudes or types of noise.
There are no visible differences between $GPML_O$, $GPML_S$ and $GPML_L$, thus, a varying length-scale parameter seems to have no effect on the predicted noise variance.
Next, it is of interest to compare the GPML residuals to the true residuals. The bars in Fig. 3.7 indicate a good similarity of the standard deviation of their respective binned residuals (blue vs. red bars), due to the good fit performance of the GPML method. On the other hand, the nonlinear regression fit of the GPML residuals (red line) shows the discrepancy towards the simulated noise level (black dashed line). Ideally, the previously added noise level could be assesesed by the (binned) prediction intervals (gray bars), which is also not the case here.
Generally, the GPML method has the limitation that it assumes a nearly constant noise level along the x-axis. The results on data with a locally varying noise level, like in this experiment approve that the GPML method does not have the ability to capture this noise. Hence, it is more suitable to apply a method that can account for locally varying noise levels, such as the HGP method.
It was found that the HGP method shows a completely different behaviour concerning the prediction intervals. The predictive standard deviation of the HGP method is able to vary along the x-axis (Fig. 3.8, left), capturing the simulated noise level more than reasonable. The pointwise standard deviations "wobble" around the previously modeled noise, a behaviour comparable to the one observed for the pointwise derivatives of the fitted curves. With an R^2 of 0.96 (Fig. 3.8, center), all three HGPs are able to capture the underlying noise level in the data.

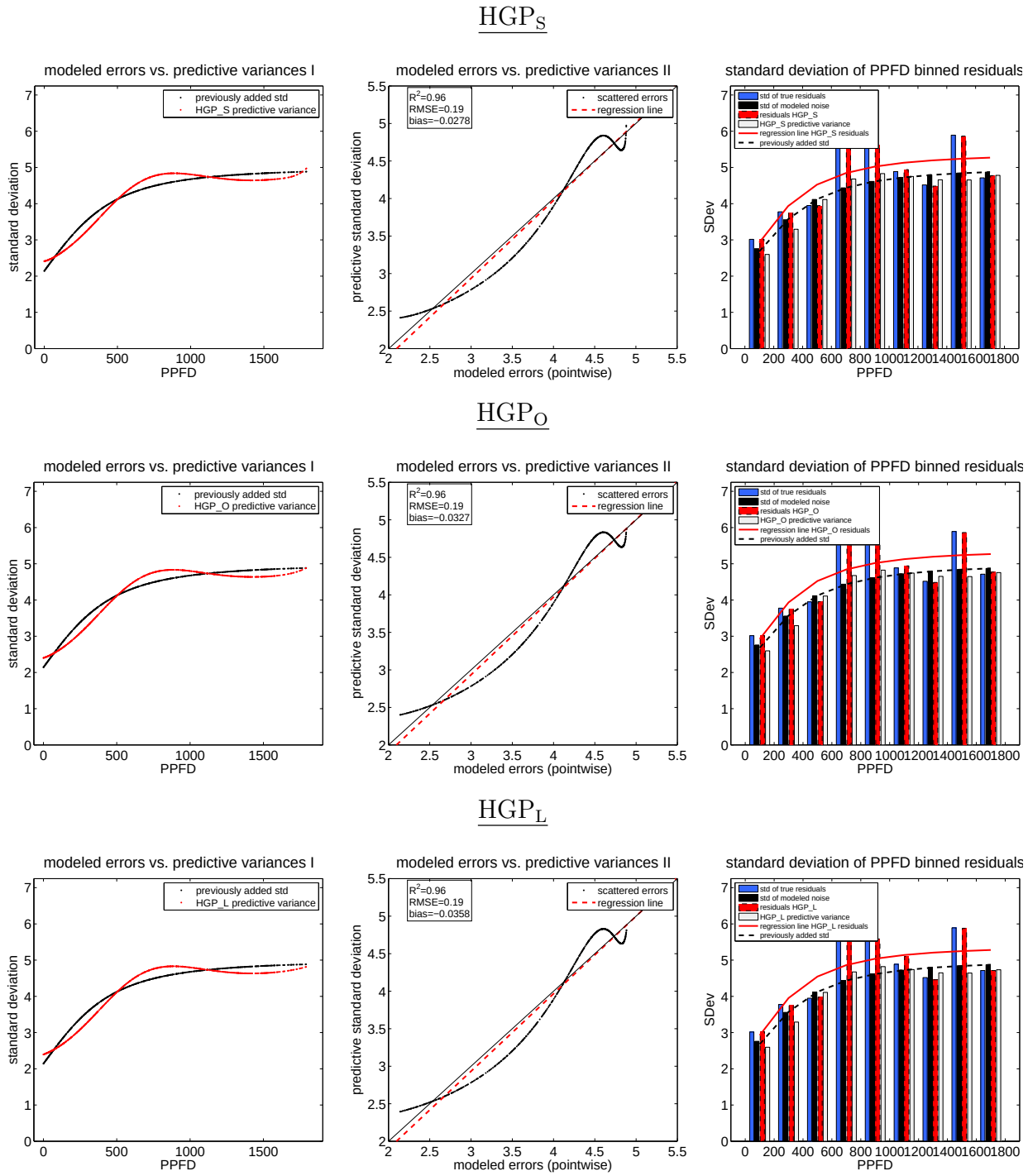


Figure 3.8: HGP predictive uncertainty analysis on the data sample shown in Fig. 3.1. For the full caption of this plot refer to Fig. 3.7.

Besides, the gray bars in Fig. 3.8 on the right prove to be a better estimator for the simulated noise level (black bars and dashed line) than the binned model residuals (red bars and line). Arguably, the predictive variance is a better estimator of the uncertainty σ than binned model residuals because the variance is estimated at every separate x by using information along the whole x-axis. In fact, the predictive variance estimated with a HGP provides a different information about the uncertainties than the binned model residuals. Binned model residuals are more dependent on the sampled noise (i.e., the true residuals depicted by the blue bars) in the respective bins and are therefore more sensitive to local noise properties in particular regions along the x-axis. They serve a different purpose by providing information about uncertainties in certain data ranges from a more local point of view. Moreover, the HGP method can provide data uncertainty estimates by the evaluation of a posterior distribution at every input point of interest in areas where data is available, which is a feature that is especially useful when it comes to applications such as uncertainty estimation for temporal aggregates (Section 3.5). Hence, a comparison to methods applied previously for that kind of application (Lasslop *et al.*, 2008) seems worthwhile.

3.3 Ecosystem Light Response Data (ELR)

After having pointed out particular strengths and weaknesses of the GP predictions on a simulated light response data set where the expected outcome is in fact already known beforehand, the methods were then run on comparable real world data. In this Section, the ecosystem light response is studied with respect to either one or two light variable(s) as a driver, built upon a measured data set described in the following.

3.3.1 Data

GP predictions were run on daytime data from three successive summers measured at the flux tower in Hainich, Germany (DE-Hai). The data set ("sample p0268") was quality checked and filtered by different criteria such as implausible or unbiological values and outlier data points (Moffat, Accepted). NEP was available as a response variable and $PPFD$, its diffuse fraction f_{dif} and the two derived variables $PPFD_{dif}$ and $PPFD_{dir}$ as input variables.
The Smith sigmoid function (eq. 3.1) is deployed as a reference function for the GP fits, since it is known to be a good representation of the light response. Hence, the parameter

set given by NLR on the data sample will be used to estimate the parameters and draw the reference function in the plots.

3.3.2 Light Response to one driver ($PPFD$)

First, both the GPML and the HGP method were run with $PPFD$ as a single input driver, investigating the $NEP(PPFD)$ relationship. The results are shown in Fig. 3.9 and are comparable to the results in the previous Section on the simulated data. Considering the fit performance, both GP methods gave an R^2 of 0.79, which equals the performance of comparable methods such as ANNs (Moffat, Accepted) on the same data. Also for runs with one of the other three variables as a single input variable, i.e. $PPFD_{dif}$, $PPFD_{dir}$ and f_{dif}, the R^2 coefficients were very similar to the ones derived from the ANN model (results not shown). Another positive result is that the Smith sigmoid function is always within the estimated GPML confidence intervals, despite the GPML drawbacks pointed out in Section 3.2.4. The regression line of the standard deviations of the binned model residuals against NEP resembles the one derived from the ANN model residuals (eqn. 3.12 or Moffat (Accepted)):

$$SDev^* = 2.47 + 0.14 \cdot NEP(PPFD). \tag{3.14}$$

Next, the binned predictive variances by an HGP on the same dataset (Fig. 3.10(b)) were analyzed, since the experiments in Section 3.2 suggest that these are more suitable to estimate uncertainties than the model residuals. The linear regression on the binned

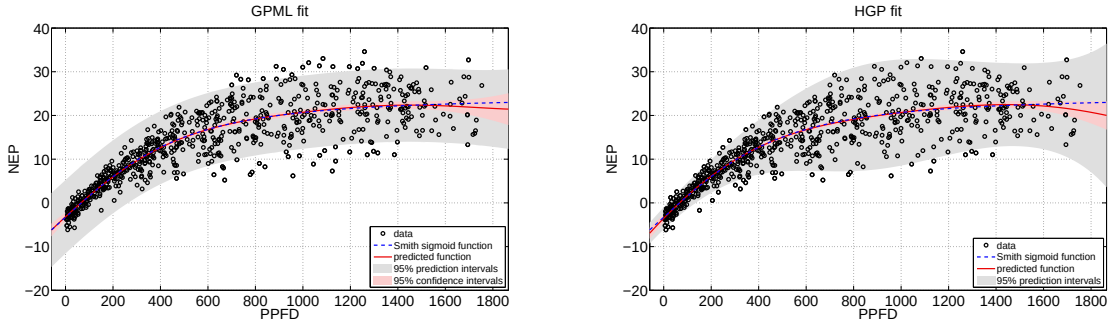


Figure 3.9: GPML and HGP prediction for the light response measured at the flux tower in Hainich, previously filtered daytime summer data from the years 2000-2002 (p0268).

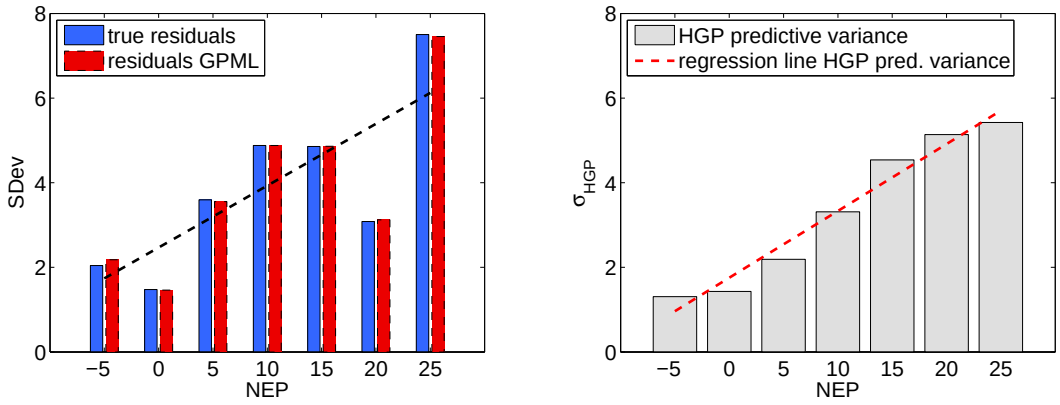


Figure 3.10: Uncertainty analysis on sample p0268: (a) Standard deviation of binned residuals. (b) Predictive variance suggested by the HGP.

predictive variances is

$$\sigma_{HGP} = 1.76 + 0.16 \cdot NEP(PPFD), \tag{3.15}$$

indicating a noise level with a comparable slope but a smaller offset. It is interesting to note, that this result is closer to the uncertainty estimates of an ANN that ran on the same dataset with 25 drivers (slope of 1.37($\pm$0.04), Moffat (Accepted)). Thus, when analyzing residuals only, the uncertainties could be overestimated, especially at low light intensity.

One more result is a steadily increasing HGP uncertainty prediction, meaning that whereas the binned residual standard deviations (Fig. 3.10(a)) are fairly distant from the noise trend for two bins, the HGP predicts a noise level that increases continuously with the flux magnitude.

3.3.3 Light Response to two drivers ($PPFD_{dir}$,$PPFD_{dif}$)

It was found by Moffat *et al.* (2010) that the diffuse radiation is the most relevant secondary control of the daytime NEP response at Hainich forest. Thus, the subsequent step is to run the GPs on data including the diffuse proportion of $PPFD$ as a second driver and compare the results with previous studies. Since total $PPFD$ is the sum of its diffuse and direct proportion, the variables $PPFD_{dir}$ and and $PPFD_{dif}$ were selected as input drivers.

In higher dimensional spaces it is useful to apply covariance functions with automatic

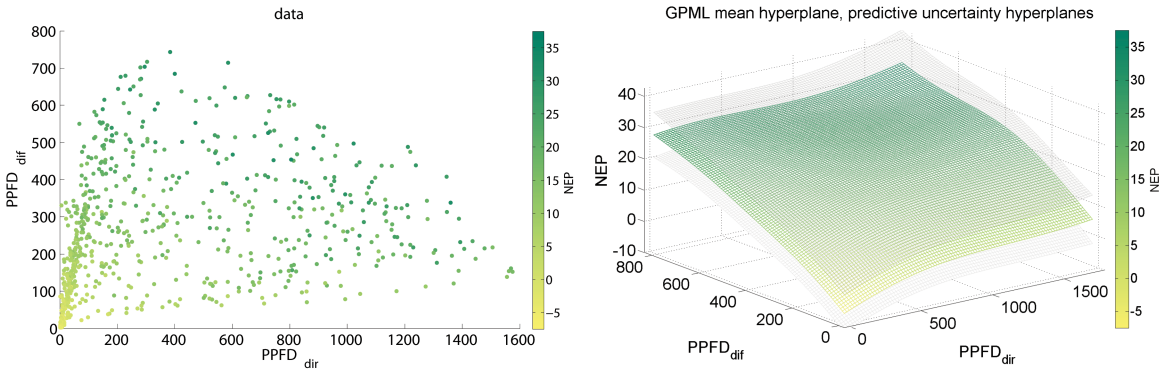


Figure 3.11: GPML prediction for the light response measured at the Hainich flux tower, previously filtered daytime summer data from the years 2000-2002 (p0268) (a) The observed data with a green color gradient indicating NEP. (b) The predicted GPML mean hyperplane. The following length-scale hyperparameters have been optimized for the input variables $[PPFD_{dir}, PPFD_{dif}]$: $\sigma_l = [2.83, 2.40]$. The gray transparent hyperplanes show the predictive uncertainties.

relevance determination (ARD), i.e. every input dimension has its own length-scale parameter.

There is a similarity of the predicted hyperplane (Fig. 3.11(b)) to the one predicted by an ANN on a comparable data sample, since it shows equal features such as a steeper initial slope for $PPFD_{dif}$ than for $PPFD_{dir}$. Also the NEP response on the diffuse light does not saturate. Moreover, it is a notable result that the length-scale parameter of $PPFD_{dif}$ is smaller than the one of $PPFD_{dir}$ ($\sigma_l = [2.83, 2.40]$), suggesting that the diffuse fraction of $PPFD$ is more important than the direct proportion of $PPFD$. These results are in agreement with previous studies, such as Moffat *et al.* (2010) and Gu *et al.* (2002). They can be finally interpreted when the results on artificial data with two input drivers are completed and included in Section 3.4. In the future, it would also be worthwhile to scatter the predictions against the predicted hyperplane of Moffat (Accepted) on the same data set and test if that hyperplane is always within the confidence intervals of the GP. When comparing the binned predictive variances (Fig. 3.12), the HGP prediction shows steadily increasing noise levels again, with a regression line of

$$\sigma_{HGP} = 1.59 + 0.1045 \cdot NEP(PPFD_{dir}, PPFD_{dif}). \tag{3.16}$$

Notably, the offset in eq. 3.16 is even smaller and thus closer to the aforementioned comparison run with 25 drivers.

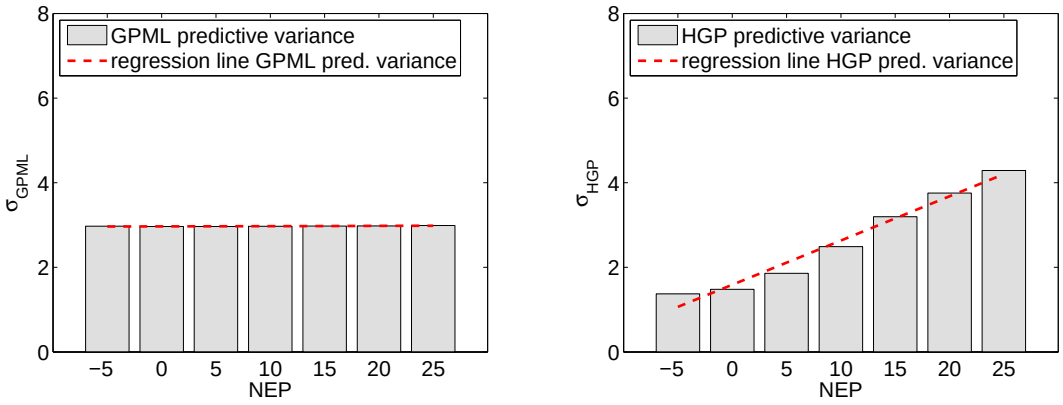


Figure 3.12: HGP Uncertainty analysis on sample p0268 with input drivers $PPFD_{dir}$ and $PPFD_{dif}$: (a) Standard deviation of binned residuals. (b) Predictive variance suggested by the HGP.

The binned predictive variances also increase if compared against either $PPFD_{dif}$ or $PPFD_{dir}$ separately (bars not shown). The predictive uncertainty planes (Fig. 3.13) suggest that the noise levels increase with an increasing flux magnitude. Fig. 3.13(b) shows that darker data points usually overlay a darker (and thus higher) HGP predictive uncertainty.

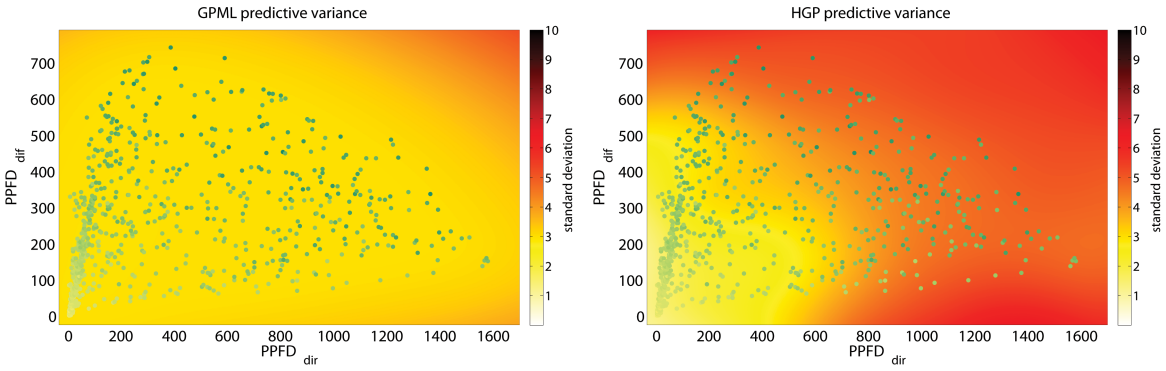


Figure 3.13: Predictive uncertainties for the light response $NEP(PPFD_{dir}, PPFD_{dif})$ measured at the Hainich flux tower. (a) σ_{GPML}. (b) σ_{HGP}. The green color gradient for the data points is the same as in the previous figures

3.4 SLR including Temperature

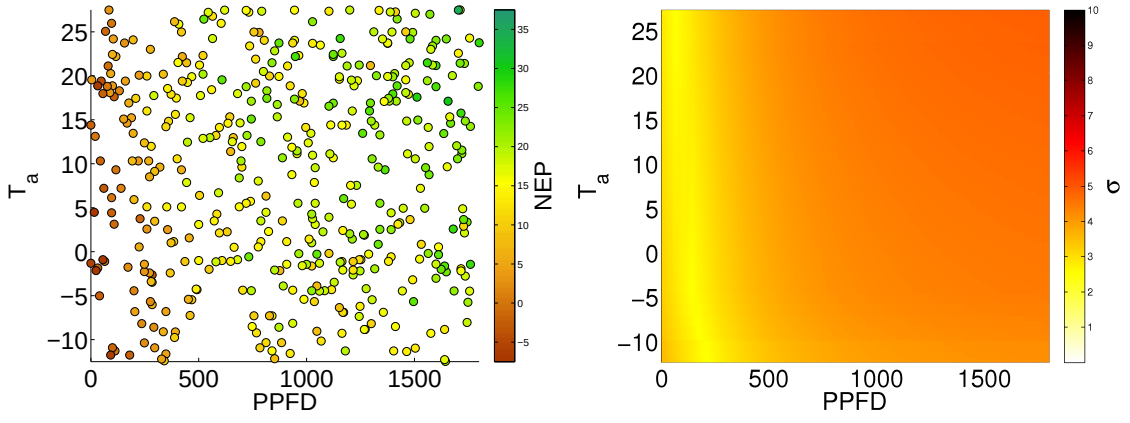
The experiments presented in this Section are the sequel to the runs in Section 3.2. The difference here is that air temperature (T_a) was added as a second input driver.

3.4.1 Data

The response of NEP to light and temperature (Fig. 3.14) was simulated by the Lloyd-Taylor model (eq. 3.10) for the temperature response and a Smith sigmoid function (eq. 3.1) as the light response, parameterized again following Moffat (Accepted):

$$NEP = \frac{\alpha \cdot GPP_{opt} \cdot PPFD}{\sqrt{GPP_{opt}^2 + (\alpha \cdot PPFD)^2}} - Rb \cdot \exp\Big(E_0 \cdot \Big(\frac{1}{T_{ref} - T_0} - \frac{1}{T_a - T_0}\Big)\Big). \quad (3.17)$$

GPP_{opt} is the gross primary production at light saturation and α the initial quantum yield, both parameters had the same value as in the 1D input experiment in Section 3.2. In the respiration model, T_0 was set to -46.02°C and the activation energy parameter E_0 to -68.15°C (Reichstein *et al.*, 2005; Lasslop *et al.*, 2010). For simplification, the parameter Rb (base respiration at reference temperature) was fixed at 4 μmol CO_2 m^{-2} s^{-1}, with the according reference temperature T_{ref} set to 15°C as in Reichstein *et al.* (2005). Heteroscedastic noise with a standard deviation σ that increases with the flux



(a) sample (b) noise

Figure 3.14: Simulated data setup for the 2D experiment. (a) Data sample. (b) Underlying (heteroscedastic) noise hyperplane calculated by eq. 3.12.

magnitude was added (Fig. 3.14(b)). The data distribution is similar to sample p0268 (measured summer data from DE-Hai years 2000-2002) as described in Section 3.2.1, resulting in a total of 451 data points (Fig. 3.14(a)). Here, the prediction was also extended to a regular data grid of 100×100 data points mapping the whole data range along the two input drivers. The grid is required to visualize the predicted function and noise hyperplanes.

3.4.2 Performance

One standard Gaussian Process (GPML) and one heteroscedastic Gaussian Proces (HGP) were optimized on the data set, the local linear regression (LOWESS) was employed as a benchmark method and a nonlinear regression (NLR) as the reference method (based on the model function in eq. 3.17). The GPML (Fig. 3.15, center) and the HGP had an R^2 performance of 73%, compared to LOWESS (Fig. 3.15, bottom) with 72% and the corresponding generating function with 71%. Thus, the fit performance was the highest possible, even slightly fitting the noise levels in some areas.
Specific weaknesses of the GPML and HGP method in capturing local features are visible at the edges again, similar as in the experiment with 1D input. This is indicated for the GPML by the overlapping planes in the right plots in Fig. 3.15, and for HGP when the predicted mean function values overlay the data generating hyperplane (Fig. 3.16(a)). By rule-of-thumb, the more the data points stand out from the underlying plane, the worse is the fit in that area. If data points are not visible at all, this indicates consistency. Note that this rule must not be overstrained, i.e. if there would be no data points visible at all this would rather indicate overfitting than a good fit. The LOWESS fit was more wobbly at medium $PPFD$ range than GPML and HGP, but more stable next to the edges. Clearly, the light saturation at high $PPFD$ is only captured by the NLR reference method (Fig. 3.15, top).
All the fitted planes show a slight gradient with increasing T_a. However, the variability originating from this driver is low, as indicated by an R^2 of 0.05 for all methods when run with T_a as a single input variable only. In contrast, if run with $PPFD$ as the only driver, 68% of the variability could be captured. The influence of T_a is more pronounced (but still at the same order of magnitude) when some parameters of the Lloyd-Taylor model vary over time (in particular Rb and E_0), which will be one of the subjects in Section 3.5.

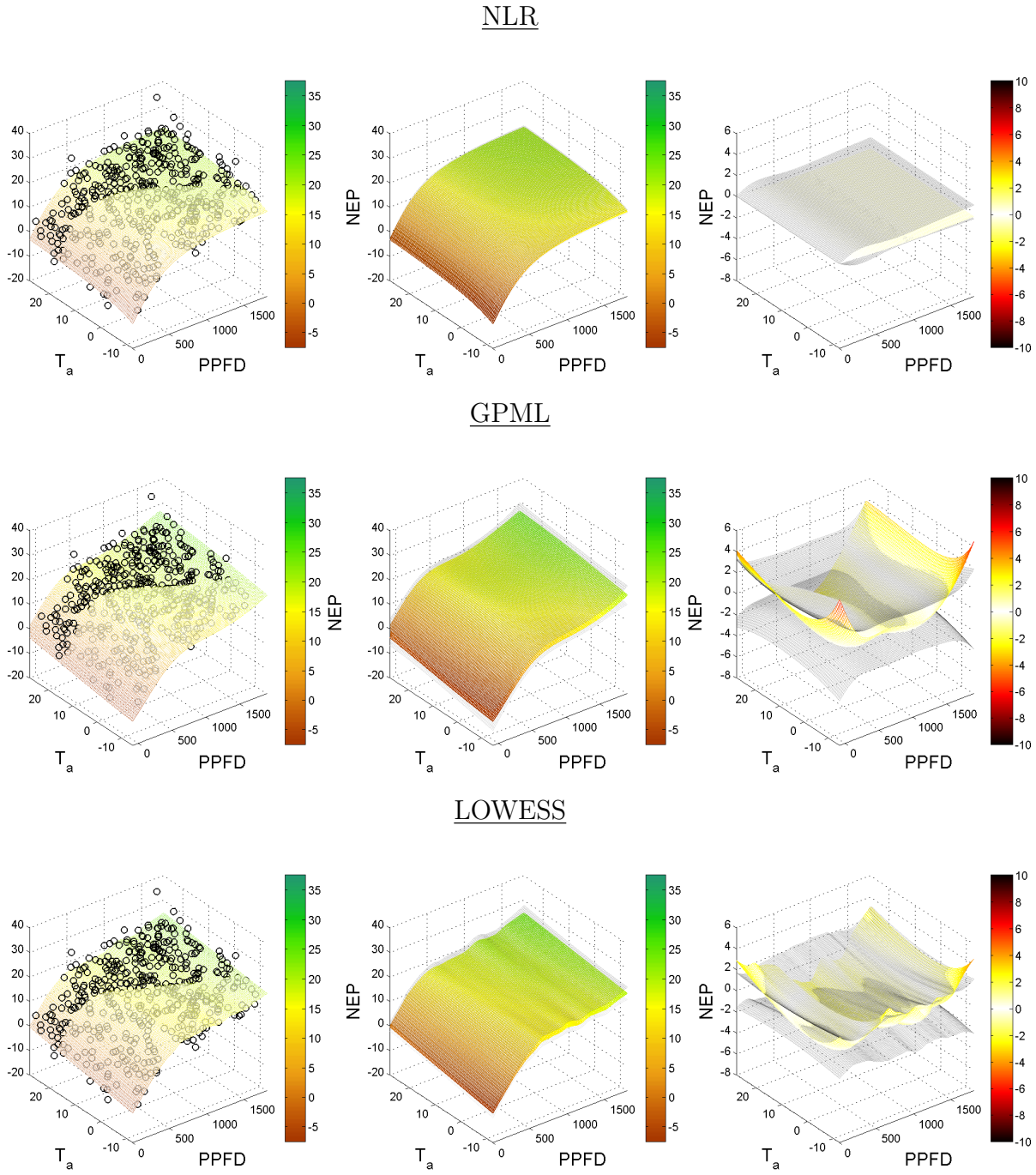
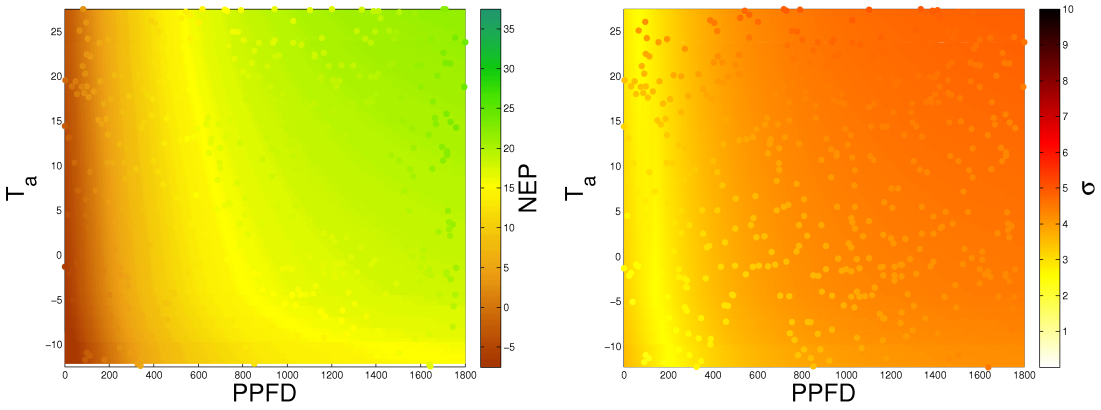


Figure 3.15: NLR, GPML and LOWESS performance on the 2D input data sample shown in Fig. 3.14.
Left: The predicted hyperplane slicing through the data sample. The color gradient from brown to green indicates NEP in μmol CO_2 m^{-2} s^{-1}. **Center:** Mean predicted function and the corresponding confidence intervals as transparent hyperplanes above and below. **Right:** Deviation of the predicted function towards the true function and the predicted confidence intervals (transparent hyperplanes). The level of deviation is quantified by a color gradient from white to red.

The confidence intervals, depicted by the gray hyperplanes in the central and right plots in (Fig. 3.15), seem to be underestimated by both GPs and LOWESS similarly as in the 1D input experiment. The deviations from the true generating function are beyond the confidence intervals particularly close to the edges of the function. It remains unclear if these overlaps are supposed to be within the 95% confidence intervals for the function or are mainly caused by the bad fit performance of the predicted mean function in these regions. Regarding the experiment here and the application experiment in Section 3.5, the focus is on the estimation of the predictive uncertainties, which is the subject of the following section.



(a) data (b) noise

Figure 3.16: Visual HGP prediction evaluation. (a) The predicted mean values at the data points overlay the simulated function hyperplane. (b) The predictive uncertainties (σ_{HGP}) at the data points overlay the simulated noise hyperplane (σ).

3.4.3 Prediction Intervals

For this 2D input experiment, the predictive uncertainties σ_{HGP} are depicted no longer as gray shaded areas as in the 1D input experiment, but either as gray transparent hyperplanes overlaying the predicted mean function (e.g., in Fig. 3.17, left), or as 2D hyperplanes with a color gradient from white to red (e.g., in Fig. 3.17, center). In the latter plot it is shown that the predicted noise levels increase with the flux magnitude, i.e. the hyperplane changes it color towards red the higher the NEP.
Another way of looking at it is to project the predicted noise levels directly on the noise generating hyperplane, as depicted in Fig.3.16(b). The less noise points visible on that

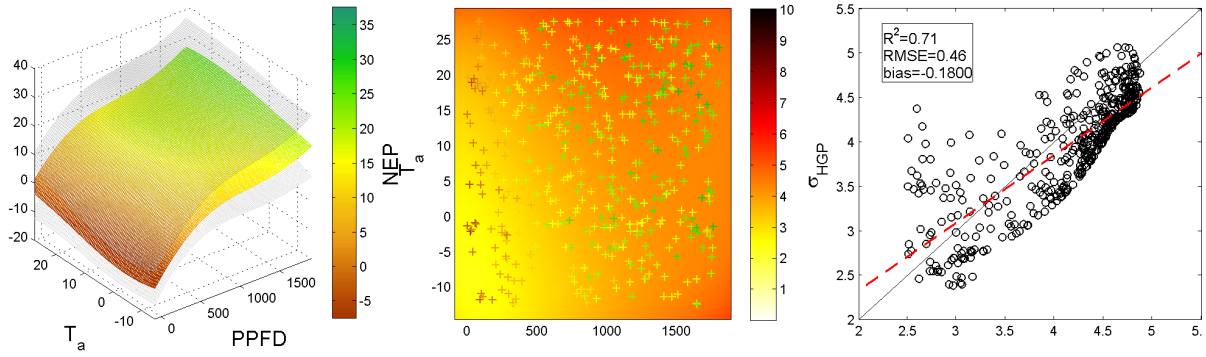


Figure 3.17: HGP fit and analysis of predictive uncertainties.
Left: The predicted mean hyperplane and the corresponding prediction intervals as gray transparent hyperplanes above and below. The color gradient from brown to green indicates NEP in μmol CO_2 m^{-2} s^{-1}. **Center:** The scattered data points overlays the predictive uncertainty hyperplane. **Right:** Scattered noise levels, HGP prediction uncertainty σHGP versus simulated σ.

hyperplane, the better the noise prediction. The deviations peak in the order of 2-3 NEP in medium $PPFD$ ranges from 400-1000 μmol photon m^{-2} s^{-1} and in T_a ranges between $-5 - 10$°C, and close to the corners of the plot.

The R^2 in the noise comparison (Fig. 3.17, right) indicates that 71% of the noise level could be captured by the HGP. Fig. 3.18 shows that the predictive variances of the HGP are a better estimator of the noise levels than binned model residuals. This result holds true when investigating the noise relationships both towards $PPFD$ and T_a. Despite some exceptions where the blue bins (model residuals) are closer to the black bins (σ) than the gray bins (HGP predictive uncertainties), the sum of the differences of the binned σ_{HGP} is always smaller than the sum of the binned model residual differences. If the uncertainties are compared against NEP (Fig. 3.18, bottom), even every single binned HGP uncertainty bar is less distant to the true noise bin than the corresponding model residual bin.

Notably, the explained variability in the noise levels (71%) is worse than the corresponding value for the 1D input experiment (96%), but still a good result given that the noise levels varies along both input dimensions. Moreover, the ten outlier points in the scatterplot (Fig. 3.17, right) with a highly overestimated σ are exclusively caused by values with very low $PPFD$ and a concurrent high T_a (Fig.3.16(b), "upper left corner").

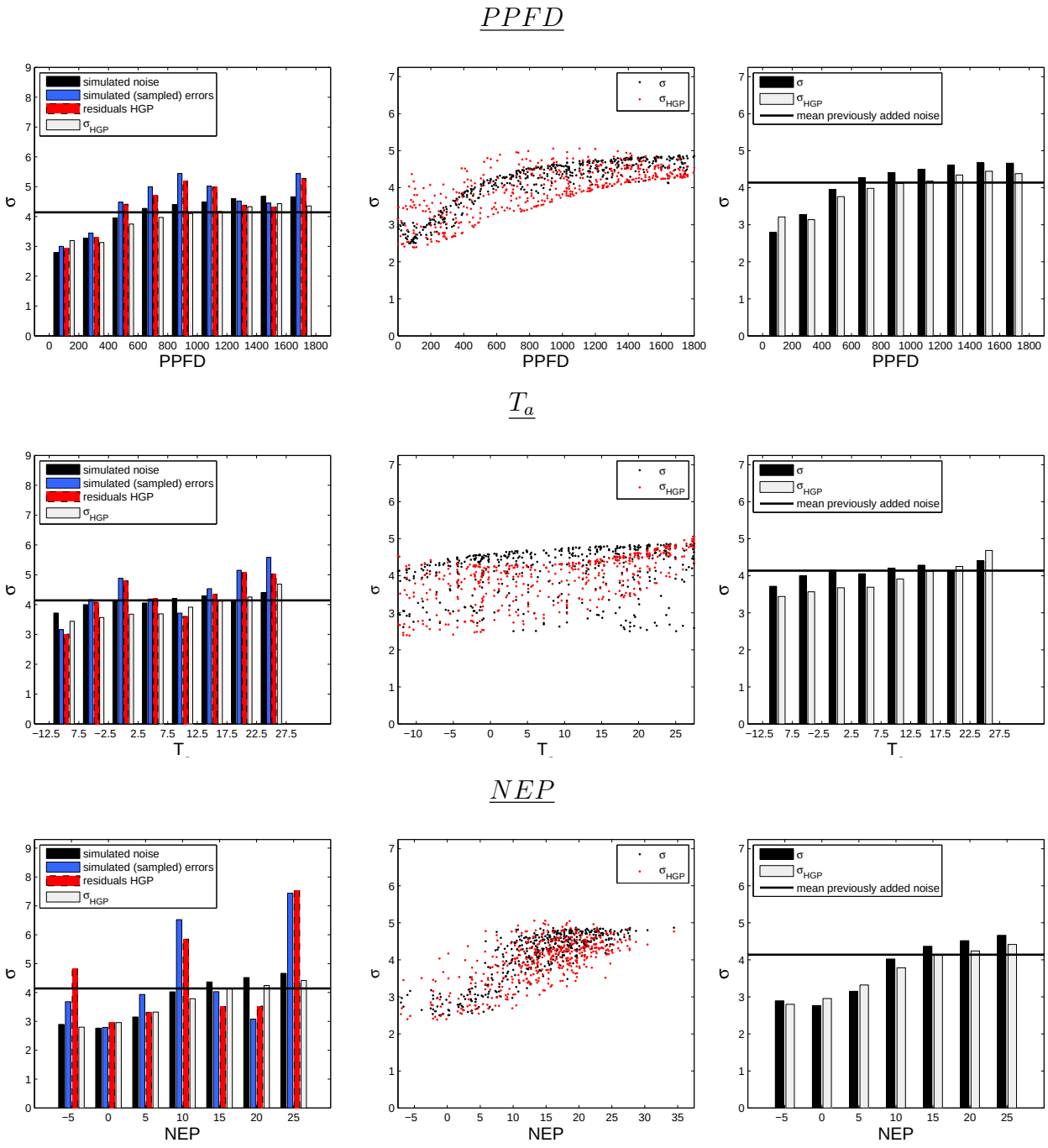


Figure 3.18: Comparison of binned residuals with HGP predictive uncertainties against $PPFD$, T_a and NEP.
Left: Binned model residuals compared to binned predictive uncertainties; and binned true residuals compared to HGP residuals. **Center:** Direct, pointwise comparison of simulated noise levels σ and predicted noise levels σ_{HGP}. **Right:** Binned predictive uncertainties against binned previously uncertainties.

3.5 ELR: Uncertainty Estimates for Annual Sums

This experiment is designed to estimate the random error in annual NEP sums, which has recently been a topic of high interest (Baldocchi, 2003; Hollinger & Richardson, 2005; Baldocchi, 2008). The possible sources of random error and the latest studies about the uncertainties associated with eddy flux measurements are described in Section 2.1.3.
The typical range of annual uncertainties in the net exchange of CO_2 reported in the literature is between 30 and 100 $\mathrm{gCm^{-2}yr^{-1}}$ (Baldocchi, 2008). When the site conditions are nearly ideal, the bound is regarded to be less than $\pm 50\,\mathrm{gCm^{-2}yr^{-1}}$ (Baldocchi, 2005).
A key result of the studies of random error in flux data is a heteroscedastic error distribution, i.e. a distribution with an inhomogeneous variance. Heteroscedastic Gaussian Processes (HGPs, Section 2.4.4) have been shown to have a good ability to estimate predictive uncertainties for such data sets, both in the literature (Kersting *et al.*, 2007; Quadrianto *et al.*, 2009) as well as in the experiments on simulated data in this work (Sections 3.2 and 3.4). This section reports on a pilot study testing and evaluating the performance of HGPs on annual ecosystem measurements.

3.5.1 Data

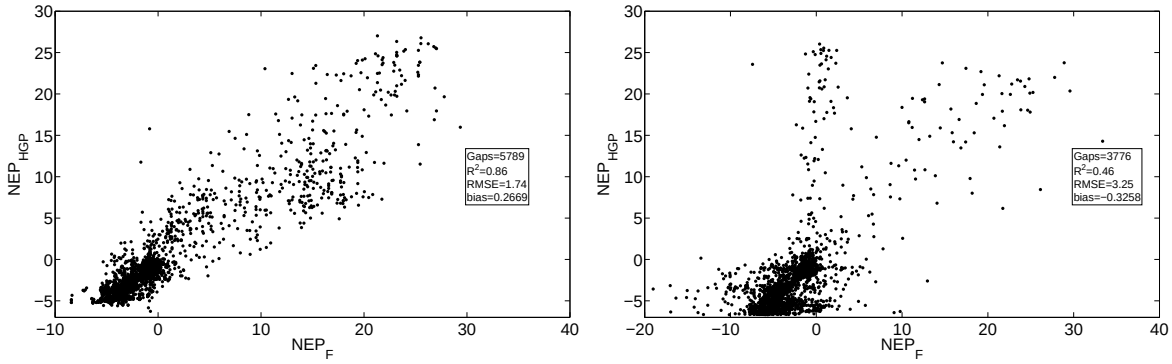
Here, the ecosystem CO_2 exchange will not only be modeled as a response to light ($PPFD$) and air temperature (T_a) as in the previous section, but also to a time variable ($time_d$) to account for the autocorrelation, which has been shown to be present both in the flux measurements and the associated random erros (Goulden *et al.*, 1996; Lasslop *et al.*, 2008). To make model predictions on highly temporal resoluted data, the benchmark data set from the gap-filling comparison study of Moffat *et al.* (2007) was used. It contains half-hourly flux measurements according to Papale *et al.* (2006) and has gap-filled meteorological variables $PPFD$ and T_a, with the gaps filled by interpolation or Artificial Neural Networks (for details refer to the Appendix of Moffat *et al.* (2007)). A complete annual meteorology is crucial for the HGP model in order to make predictions over the whole year. Unplausible $PPFD$ values smaller than zero, as they occur often during nighttime, were set to zero.
Measurements from the Hainich forest (DE-Hai, cf. Section 2.1.2) were picked to be able to compare the error models to the results in Section 3.3. Also, another site in Hesse, France (FR-Hes), was chosen, because with 78% it had the highest annual data avail-

ability of all sites of the benchmark set (90% during daytime, 43% during nighttime). Since it is a forest dominated by *Fagus sylvatica*, same as the Hainich forest, similarities might be detectable. Data from the year 2001 was picked for both sites.
The first runs were performed on training data sets with the maximum available amount of data, i.e. $n = 13744$ for Hesse and $n = 11731$ for Hainich. However, runtime issues caused by the necessary matrix inversion which is bound by $O(n^3)$ were found to be inacceptable, especially with regard to maintain a realistic possibility to apply the HGP runs to other sites. The runtime for the above sample sizes were up to one week, not only caused by the matrix inversions, but also by the optimization runs for the HGPs. Consequently, the sample size was reduced to $n/3$, by selecting only every third half-hourly measurement, independent of occuring gaps. Two subsamples were taken, the first starting at the first half-hour of the year (denoted as HGP I from here on), and the second subsample starting at the second half-hour of the year (HGP II). Since these subsamples start at different indices, they do not share any data point at all (i.e., they are *disjunct*). In turn, the resulting runtimes with samples of dimension $n/3$ were reduced to one day or less. Moreover, this approach offers the possibility of comparing the two runs and evaluating the robustness of the HGP method; also, averages over the two models can be calculated (HGP*). The approach can easily be extended to a third subsample for each siteyear, which is due to time constraints not part of this thesis.

3.5.2 Annual Aggregates

The two trained HGPs were employed for prediction over the whole annual time series on a half-hourly basis, on a previously gap-filled data set. To evaluate the predicted annual NEP course predicted by the HGP, it was compared to the measured NEP with gap-filled values following Reichstein *et al.* (2005), which is denoted as NEP_F. Moreover, the NEP predicted with a hyperbolic light response curve (NEP HBLR) by Lasslop *et al.* (2010) was available.
A quality check of the filled annual sums was performed by comparing the HGP gap-filled NEP values to the results of Reichstein *et al.* (2005). The R^2 of 87% for DE-Hai indicates good agreement between the two modeling approaches (Fig. 3.19(a)), and thus, a reasonable HGP annual sum. The gap comparison in FR-Hes shows some outliers with an apparent overestimation for NEP_{HGP} (in the range [-2,2] NEP_F), which are exclusively resulting from 4 summer days with $T_a > 15$ and average $PPFD > 800$. This could be an artifact of the HGP prediction or an indicator for measurement problems in either the fluxes or the meteorology around that time. However, nearly identical outliers

were observed independent from the training data subsample.
The results of the annual NEP sums are summarized in Table 3.2. For both sites, the HGP runs on the two samples resulted in very similar annual aggregates, each within the 95% prediction intervals of one another. The HGP* annual NEP sum of DE-Hai is in a comparable magnitude to the HBLR prediction, whereas the pendant for FR-Hes is rather similar to NEP_F.
The averaged NEP along the two model runs, HGP* should be the more plausible option to use for reporting sums and uncertainties compared to single runs, which is for two reasons. First, the marginal likelihood is a relative measure, and can thus not be compared between the two different samples in order to select the supposedly better model. Second, by averaging over the two models HGP I and HGP II, more information is included into the prediction.



(a) DE-Hai (b) FR-Hes

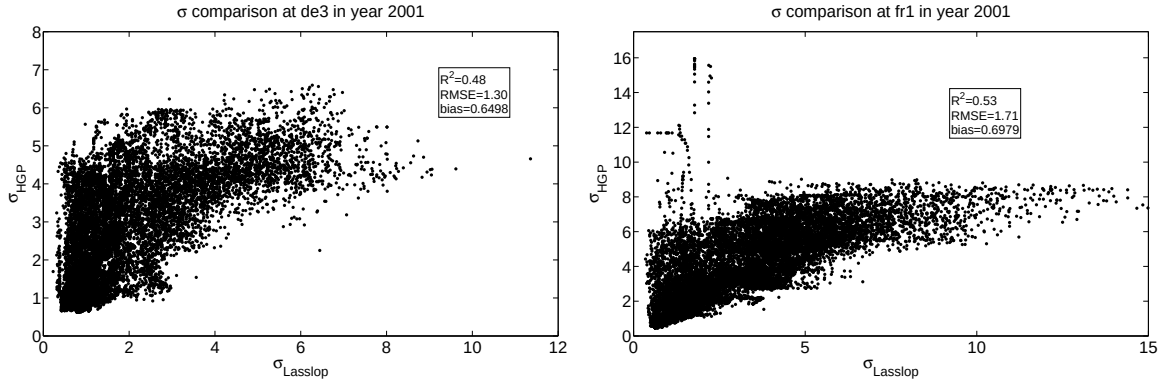
Figure 3.19: Gap-filling comparison: HGP II vs. Reichstein *et al.* (2005)

	DE-Hai 2001	**FR-Hes 2001**
NEP HGP I	520.53 (±14.35)	580.17 (±24.76)
NEP HGP II	531.19 (±13.83)	593.02 (±22.28)
NEP **HGP***	**525.86 (±14.09)**	**586.60 (±23.52)**
NEP_F	608.13 (±12.99)	599.21 (±21.52)
NEP HBLR	511.18	574.91

Table 3.2: Annual aggregates of NEP [gCm^{-2}yr^{-1}] for the sites in Hainich and Hesse. 95% annual uncertainties are given in brackets by the 1.96-fold prediction interval (NEP HGP I and HGP II) and the 1.96-fold standard deviation (NEP_F).

3.5.3 Annual Predictive Uncertainties

The 95%-prediction intervals calculated from the posterior of the HGP have been used to estimate the random error in the annual NEP sums. The 17520 uncertainty values for each site have been aggregated by a standard Gaussian error propagation.
The good quality of the uncertainty estimates for the annual sums is plausible, because they make the annual sums overlap with one of the two reference methods for both sites(Table 3.2). Moreover, the order of magnitude of the aggregated random errors is the same as reported in Lasslop *et al.* (2008). In this approach, random error estimates ($\sigma_{Lasslop}$) were derived from the gap-filling algorithm of Reichstein *et al.* (2005), computing the expected value of the flux using data measured under the same meteorological conditions in a time window of ± 7 days.
Differences become apparent when directly scattering the two approaches against each other (Fig. 3.20). Whereas σ_{HGP} shows slight signs of saturation in the range of 6-8 NEP, σ_{HGP} does seemingly not saturate. The outliers at site FR-Hes for $\sigma_{HGP} > 8$ are exclusively from winter data between day 27 and 37 and could either be an HGP artifact or indicate problems with the measurements.
Analyzing the binned predictive variances σ_{HGP} as in the earlier Sections of this Chapter, typical error properties of CO_2 flux measurements are revealed again (Fig. 3.21). First, the noise level is increasing with the flux magnitude (heteroscedasticity), it is steadily increasing, as shown when the binned variances are plotted against NEP. This is in agreement with the results in the literature (Lasslop *et al.*, 2008; Richardson *et al.*,



(a) Hainich (b) Hesse

Figure 3.20: Noise comparison: σ_{HGP} vs. $\sigma_{Lasslop}$. The latter is reported in the study of Lasslop *et al.* (2008).

2006, 2008). FR-Hes shows a considerably higher uncertainty bar for -5 NEP than for 0 NEP, which is a result especially supporting the idea of errors increasing with the flux magnitude. For the site DE-Hai, there is also large agreement with the corresponding studies in Section 3.3, where uncertainties in NEP were derived from previously filtered summer daytime data.

Similar noise patterns between the two sites were found, which become evident when comparing σ_{HGP} binned against $PPFD$ and T_a, even solely by visual inspection. The magnitude of the random error estimates is higher for FR-Hes constantly for all bins in the three columns for binning against $PPFD$, T_a and NEP, which might be due to more unfavorable site conditions than at the Hainich site.

Different theories exist about whether the random errors in flux measurements are Gaussian or Laplacian distributed (Section 2.1.3). An analysis of the distribution of the predicted noise levels here would not provide surprising information, since Gaussian Processes are constraint to the assumption of Gaussian observation noise. However, it is an interesting finding that characteristic flux data noise properties can be captured when assuming the random errors to be Gaussian.

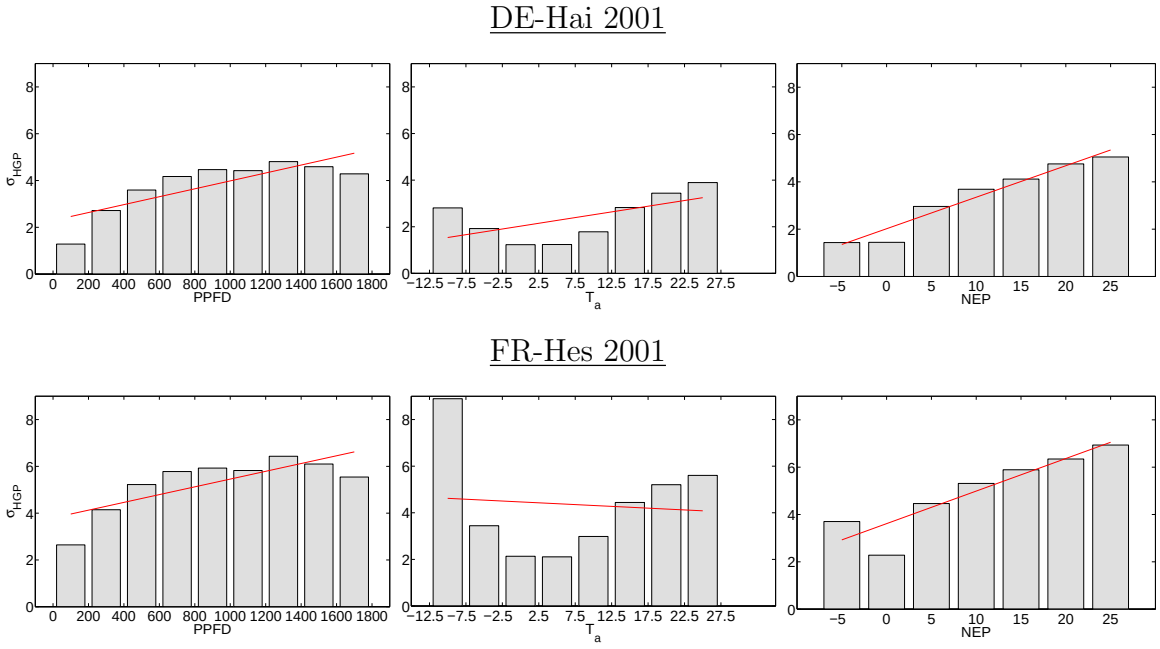
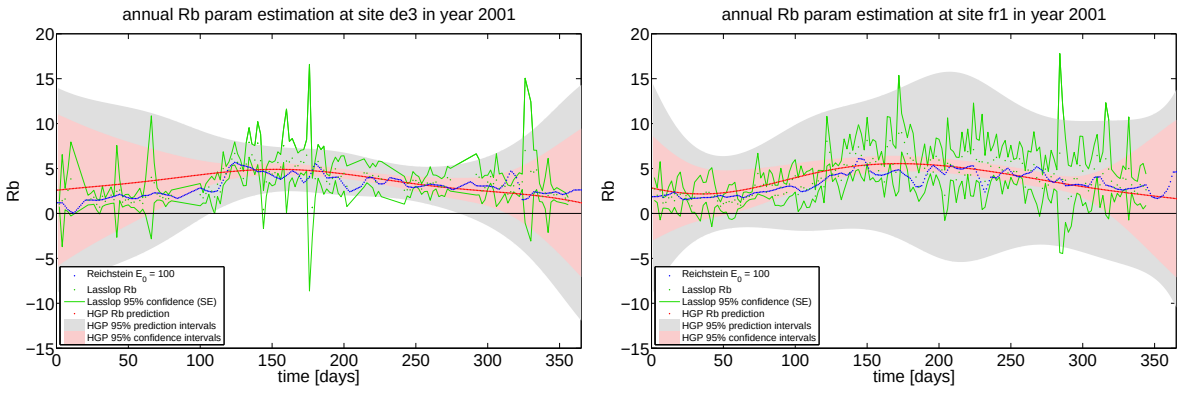


Figure 3.21: Predictive uncertainty analysis for DE-Hai and FR-Hes, year 2001. The binned predictive variances σ_{HGP} are plotted against $PPFD$ (Left), T_a (Center) and NEP (Right)

3.5.4 Flux partitioning

For flux partitioning of NEP into its components GPP and R_{eco}, the latter was modeled by the Lloyd-Taylor model (eq. 3.10, Lloyd & Taylor (1994)). The estimation of the Rb parameter (base respiration at reference temperature) of this model was performed by the algorithm described in Section 3.1.3, with the subsequent steps of partitioning the fluxes. The results for the Rb estimation can be compared with two different benchmarks. First, with the study of Reichstein *et al.* (2005), providing continuous time series of R_{eco} for $E_0 = 100$. With this information, Rb could be calculated analytically from the Lloyd-Taylor model. Second, the flux partitioning study of Lasslop *et al.* (2010) provided Rb estimates and associated standard errors (SE).
Fig. 3.22 shows a general agreement between the HGP predictions and the two reference partitioning models, as indicated by the fraction of corresponding data points within the HGP confidence intervals. On a daily basis, for DE-Hai, 4.8% (HGP I) and 3.9% (HGP II) of the Rb estimates of Reichstein *et al.* (2005) are beyond the HGP 95% confidence intervals, for FR-Hes: 5.9% (HGP I) and 4.7% (HGP II). However, the HGP predictions seemingly do not capture the full variability of the Rb parameter. It is relatively well estimated only in average. If one would be interested in more precise Rb parameter estimates with an HGP or GPML model, more sophisticated covariance functions accounting e.g. for the diurnal cycle or a periodicity in the data might be a solution. Two more interesting observations can be inferred here. First, the confidence intervals of the HGP prediction are considerably higher during winter than during summer. An



(a) DE-Hai (b) FR-Hes

Figure 3.22: Annual Rb estimation with uncertainties on the subsamples HGP I. For a legend, refer to the box inside the axis.

interesting next step would be to include a fixed number of preceding and following days (e.g., 50) from the neighbouring years into the prediction and then compare the confidence intervals, but also the prediction intervals. This is, however, beyond the scope of this diploma thesis. Second, the prediction intervals in summer are by far lower for Hainich than for Hesse, which suggests a lower random error in the measurements during night at the Hainich site compared to Hesse. This also contributes to the lower annual uncertainty in the NEP sum for Hainich.
This approach was used to test the HGP ability to infer a respiration model parameter directly from the data. The Rb parameter is neither required to predict NEP, nor the corresponding uncertainties. Nevertheless, it adds further information regarding the quality of the model, e.g. implausible models could be detected in the case of very unlikely partitioned annual sums of GPP and R_{eco}. The comparison of the inferred annual aggregates of the partitioned fluxes are summarized in Table 3.3.

	DE-Hai 2001	**FR-Hes 2001**
R_{eco} HGP I	1104.42	1379.73
R_{eco} HGP II	1137.40	1291.06
R_{eco} HGP*	1120.91	1335.39
R_{eco} Reichstein	979.28	1274.75
R_{eco} Lasslop	995.61	1348.06
GPP HGP I	1624.96	1959.91
GPP HGP II	1668.60	1884.09
GPP HGP*	1646.78	1922.01
GPP Reichstein	1587.41	1873.96
GPP Lasslop	1506.79	1922.98

Table 3.3: Annual aggregates of GPP [gCm^{-2}yr^{-1}] and R_{eco} [gCm^{-2}yr^{-1}] for the sites in Hainich and Hesse. R_{eco} / GPP Reichstein following Reichstein *et al.* (2005), R_{eco} / GPP Lasslop following Lasslop *et al.* (2010),

4 Conclusion and Outlook

The main objective of this methodological thesis was to estimate uncertainties in ecosystem data with Gaussian Processes, a modern Machine Learning method. The key challenges were the nonlinear and complex character of the underlying biosphere-atmosphere interactions and the properties of noisy, multidimensional and fragmented ecosystem flux measurements (Moffat, Accepted). Gaussian Processes offer the necessary modeling power to approach these questions under relatively few assumptions and without a prescribed mathematical function, directly from a *data perspective*.

To provide insight into how Gaussian Processes work and to evaluate their benefits and drawbacks, the experiments ranged from simple, artificial scenarios to more demanding, real-world data sets:

1. Simulated data with additive noise, imitating a $NEP(PPFD)$ light response.

2. Previously filtered, summer, daytime light response data from the Hainich forest.

3. Simulated data with additive noise, imitating a $NEP(PPFD, T_a)$ light response

4. Half-hourly, annual, but incomplete time series from the Hainich forest and Hesse forest. $PPFD$, T_a and information about time were employed as input drivers.

The main results of the first, most comprehensive artificial data experiment (**1.**) revealed that GP confidence intervals tend to be underestimated, whereas prediction intervals for assessing data uncertainties are the most outstanding ability of the Heteroscedastic Gaussian Process (HGP) approach. Another finding was that weaknesses in capturing local features of a function contrasted good overall fits.
In the subsequent experiment (**2.**) on the according measured summer daytime data from the Hainich forest, the HGP predictive uncertainties suggested that random errors in NEP are overestimated if assessed with a traditional method (binning model residuals). This finding could be approved for the residuals of five different modeling approaches (LOWESS, NLR, GPML, HGP and ANN), indicating that model residuals

should rather be used for quantifying the noise properties in more local regions of a function. Also, a trained GP model suggested agreement with the existing theory that the diffuse fraction of $PPFD$ is a more relevant driver of the light response than the direct proportion of $PPFD$. In the future, it would be good to confirm these results with an analysis from data at other flux towers.
The **3.** experiment was built on the first runs, with T_a added as a second input driver and it could confirm the initial results: the noise levels in the artificial data set were still well captured, despite some expected performance losses. Here, the noise simulation was varied along 2 input dimensions, therefore forming a basis for applying HGPs in higher dimensional data sets.
The final experiment (**4.**) aimed at estimating uncertainties in annual sums of fragmented carbon flux measurements. After two HGPs were trained on two disjunct samples, aggregates could be calculated using the according, previously gap-filled meteorological time series for prediction. Estimates of annual uncertainties in NEP in the year 2001 of ± 14.1 $\mathrm{gCm^{-2}yr^{-1}}$ for Hainich and ± 23.5 $\mathrm{gCm^{-2}yr^{-1}}$ for Hesse were calculated. These are similar to the reported random error quantifications in the literature. The GP assumption of Gaussian observation noise is a limiting, but reasonable assumption, as indicated by the predicted noise patterns, which are both consistent between sites as well as with existing studies (Lasslop *et al.*, 2008; Richardson *et al.*, 2008).

The modern Machine Learning method of Gaussian Process method was shown to be a suitable framework for nonlinear regression in various experimental set ups, revealing very particular strengths and drawbacks. The GP approach is recommended to be applied in situations of Gaussian observation noise and if the sample size n does not largely exceed 10000 data points.
In the future, it will be worthwhile to estimate uncertainties in annual NEP sums with HGPs for more measurement sites of the FLUXNET observation network. Supporting work in this direction does include further artificial experiments on data sets with simulated pink noise, i.e. a simulated autocorrelation in the noise levels. Moreover, the results obtained in this work suggest that there is room for improved strategies in training HGPs for annual aggregates with e.g., 50 surrounding days of each the preceding and the following year.

Bibliography

Allen D (1974) Relationship between Variable Selection and Data Augmentation and a Method for Prediction. *Technometrics*, **16**, 125–127.

Aubinet M, Grelle A, Ibrom A, *et al.* (2000) Estimates of the annual net carbon and water exchange of forests: The EUROFLUX methodology. *Advances in Ecological Research*, **30**, 113–175.

Baldocchi D (1997) Measuring and modelling carbon dioxide and water vapour exchange over a temperate broad-leaved forest during the 1995 summer drought. *Plant Cell Environment*, **20**, 1108–1122.

Baldocchi D (2003) Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, **9**, 479–492.

Baldocchi D (2005) Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, **9**, 479–492.

Baldocchi D (2008) Breathing of the terrestrial biosphere: lessons learned from a global network of carbon dioxide flux measurement systems. *Australian Journal of Botany*, **56**, 1–26.

Baldocchi D, Falge E, Gu L, *et al.* (2001) FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, **82**, 2415–2434.

Bates D, Watts D (1988) *Nonlinear Regression Analysis and its Applications*. Wiley, N.Y.

Beer C, Reichstein M, Tomelleri E, *et al.* (2010) Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate. *Science*, **329**, 834–838.

Bishop C (2006) *Pattern Recognition and Machine Learning*. Springer.

Bonan G (2002) *Ecological climatology: concepts and applications*. Cambridge University Press, New York, NY, USA.

Bovscheverov V, Voronov V (1960) Akustitscheskii fljuger (acoustic rotor). *Izv AN SSSR, ser Geofiz*, **6**, 882–885.

Burba G, Anderson D (2010) *A Brief Practical Guide to Eddy Covariance Flux Measurements: Principles and Workflow Examples for Scientific and Industrial Applications*. ISBN: 978-0615430133. LI-COR Biosciences, Lincoln, USA, Hardbound and Softbound Editions, 211 pp.

Burden R, Faires J (2000) *Numerical Analysis*. Brooks/Cole, 7 edn.

Cleveland W (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Dempster A, Laird N, Rubin D (1977) Maximum Likelihood from incomplete data via EM Algorithm. *Journal of the Royal Statistical Society Series B-Methodologica*, **39**, 1–38.

Draper N, Smith H (1998) *Applied Regression Analysis*. Wiley Series in Probability and Statistics, 3rd edn.

Drignei D, Forest C, Nychka D (2008) Parameter estimation for computationally intensive nonlinear regression with an application to climate modeling. *Parameter estimation for computationally intensive nonlinear regression with an application to climate modeling*, **2**, 1217–1230.

Duggleby R (1995) Analysis of enzyme progress curves by nonlinear regression. *Methods in Enzymology*, **249**, 61–90.

Falge E, Baldocchi D, Olson R, *et al.* (2001) Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agricultural and Forest Meteorology*, **107**, 43–69.

Foken T (2003) *Angewandte Meteorologie*. Springer Verlag, Berlin Heidelberg New York.

Freedman D (2009) *Statistical Models, Theory and Practice, Revised Edition*. Cambridge University Press.

Gao J (2004) Modelling long-range-dependent Gaussian processes with application in continuous-time financial models. *Journal of Applied Probability*, pp. 467–82.

Goeckede M, Markkanen T, Hasager CB, Foken T (2006) Update of a footprint-based approach for the characterisation of complex measurement sites. *Boundary-Layer Meteorology*, **118**, 635–655.

Goldberg P, Williams C, Bishop C (1998) Regression with input-dependent noise: A Gaussian process treatment. In: *Advances in neural information processing systems 10*, vol. 10 of *Advances in neural information processing systems* (ed. Jordan, MI and Kearns, MJ and Solla, SA), pp. 493–499. 11th Annual Conference on Neural Information Processing Systems (NIPS), DENVER, CO, DEC 01-06, 1997.

Goulden M, Munger J, Fan S, Daube B, Wofsy S (1996) Measurements of carbon sequestration by long-term eddy covariance: Methods and a critical evaluation of accuracy. *Global Change Biology*, **2**, 169–182.

Gu L, Baldocchi D, Verma S, Black T, Vesala T, Falge E, Dowty P (2002) Advantages of diffuse radiation for terrestrial ecosystem productivity. *Journal of Geophysical Research - Atmosphere*, **107**, D06–4050.

Hollinger D, Aber J, Dail B, *et al.* (2004) Spatial and temporal variability in forest-atmosphere CO(2) exchange. *Global Change Biology*, **10**, 1689–1706.

Hollinger D, Richardson A (2005) Uncertainty in eddy covariance measurements and its application to physiological models. *Tree Physiology*, **25**, 873–885. International Conference on Modeling Forest Production, BOKU Univ Nat Resources & Appl Life Sci, Vienna, AUSTRIA, APR 19-21, 2004.

Kersting K, Plagemann C, Pfaff P, Burgard W, eds. (2007) *Most Likely Heteroscedastic Gaussian Process Regression*, Proceedings of the 24th International Conference on Machine Learning.

Knohl A (2003) *Carbon dioxide exchange and isotopic signature (13C) of an unmanaged 250 year-old deciduous forest, Chapter 2: Methods*. Doctoral thesis, Friedrich Schiller University Jena.

Kutsch WL, Kolle O, Rebmann C, Knohl A, Ziegler W, Schulze ED (2008) Advection and resulting CO2 exchange uncertainty in a tall forest in central Germany. *Ecological Applications*, **18**, 1391–1405.

Lasslop G, Reichstein M, Kattge J, Papale D (2008) Influences of observation errors in eddy flux data on inverse model parameter estimation. *Biogeosciences*, **5**, 1311–1324.

Lasslop G, Reichstein M, Papale D, *et al.* (2010) Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. *Global Change Biology*, **16**, 187–208.

Lauritzen S (1981) Time series Analysis in 1880 - a discussion of contributions made by Thiele,T.N. *International Statistical Review*, **49**, 319–331.

Levenberg K (1944) A method for the solution of certain non-linear problems in least squares. *The Quarterly of Applied Mathematics*, **2**, 164–168.

Lloyd A, Taylor J (1994) On the temperature dependence of soil respiration. *Functional Ecology*, **8**, 315–323.

MacKay D (1998) *Neural Networks and Machine Learning*, chap. Introduction to Gaussian Processes, pp. 133–165. Springer Verlag.

Macke JH, Gerwinn S, White LE, Kaschube M, Bethge M (2011) Gaussian process methods for estimating cortical maps. *Neuroimage*, **56**, 570–581.

Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, **11**, 431–441.

Matheron G (1963) Principles of geostatistics. *Economic Geology*, **58**, 1246–1266.

Müller HG, Chiou JM, Leng X (2008) Inferring gene expression dynamics via functional regression analysis. *BMC Bioinformatics*, **9**, 60.

Moffat AM (Accepted) *A new methodology to interpret high resolution measurements of net carbon fluxes between terrestrial ecosystems and the atmosphere, Chapter 7: Assessing competing semi-empirical equations*. Doctoral thesis, Friedrich Schiller University Jena.

Moffat AM, Beckstein C, Churkina G, Mund M, Heimann M (2010) Characterization of ecosystem responses to climatic controls using artificial neural networks. *Global Change Biology*, **16**, 2737–2749.

Moffat AM, Papale D, Reichstein M, *et al.* (2007) Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, **147**, 209–232.

Moncrieff J, Malhi Y, Leuning R (1996) The propagation of errors in long-term measurements of land-atmosphere fluxes of carbon and water. *Global Change Biology*, **2**, 231–240.

Montgomery R (1948) Vertical Eddy Flux of Heat in the Atmosphere. *Journal of Meteorology*, **5**, 265–274.

Neal R (1997) Monte carlo implementation of gaussian process models for bayesian regression and classification. Crg-tr-97-2, Department of Computer Science, University of Toronto.

O'Hagan A (1978) Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological),*, **40**, 1–42.

Papale D, Reichstein M, Aubinet M, *et al.* (2006) Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation. *Biogeosciences*, **3**, 571–583.

Quadrianto N, Kersting K, Reid M, Caetano T, Buntine W (2009) Kernel conditional quantile estimation via reduction revisited. In: *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining (ICDM 2009)* (eds. Wei Wang, Kargupta H, Ranka S, Yu P, Xindong Wu), pp. 938–943.

Rasmussen C (1996) *Evaluation of Gaussian Processes and Other Methods for Non-linear Regression.* PhD thesis, Department of Computer Science, University of Toronto.

Rasmussen C, Nickisch H (2010) Gaussian processes for machine learning toolbox. *Journal of Machine Learning Research*, **11**, 3011–3015.

Rasmussen C, Williams C (2006) *Gaussian Processes for Machine Learning.* MIT Press.

Raupach M, Rayner P, Barrett D, *et al.* (2005) Model-data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications. *Global Change Biology*, **11**, 378–397.

Reichstein M, Falge E, Baldocchi D, *et al.* (2005) On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*, **11**, 1424–1439.

Richardson A, Hollinger D, Burba G, *et al.* (2006) A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. *Agricultural and Forest Meteorology*, **136**, 1–18.

Richardson AD, Mahecha MD, Falge E, *et al.* (2008) Statistical properties of random CO2 flux measurement uncertainty inferred from model residuals. *Agricultural and Forest Meteorology*, **148**, 38–50.

Schölkopf B, Smola A, Williamson R, Bartlett P (2000) New support vector algorithms. *Neural Computation*, **12**, 1207–1245.

Schmid H (1994) Source Areas for Scalars and Scalar Fluxes. *Boundary-Layer Meteorology*, **67**, 293–318.

Schulze E, Wirth C, Heimann M (2000) Climate change - managing forests after Kyoto. *Science*, **289**, 2058–2059.

Siegenthaler U, Stocker T, Monnin E, *et al.* (2005) Stable carbon cycle-climate relationship during the late Pleistocene. *Science*, **310**, 1313–1317.

Smith E (1938) Limiting factors in photosynthesis: Light and carbon dioxide. *The Journal of General Physiology*, **22**, 21–35.

Storch H, Zwiers F (1999) *Statistical Analysis in Climate Research.* Cambridge University Press.

Sun Y, Brown MB, Prapopoulou M, Adams R, Davey N, Moss GP (2010) The application of Gaussian processes in the prediction of absorption across mammalian skin and synthetic membranes. *Journal of Pharmacy and Pharmacology*, **62**, 803.

US Department of Energy (2008) Carbon cycling and biosequestration. In: *Report from the March 2008 Workshop, DOE/SC-108*, pp. 2–3. U.S. Department of Energy. Office of Science.

Valentini R, Matteucci G, Dolman A, *et al.* (2000) Respiration as the main determinant of carbon balance in European forests. *Nature*, **404**, 861–865.

Williams C, Rasmussen C (1996) Gaussian processes for regression. In: *Advances in Neural Information Processing Systems 8: Proceedings of the 1995 Conference*, vol. 8 of *Advances in Neural Information Processing Systems* (ed. Touretzky, DS and Mozer, MC and Hasselmo, ME), pp. 514–520. 9th Annual Conference on Neural Information Processing Systems (NIPS), DENVER, CO, NOV 27-30, 1995.

Appendix

Software

This study and the associated experiments are exclusively implemented in `MATLAB 7.10`. The following toolboxes have been included:

GPML toolbox: Gaussian Processes for Machine Learning. A very comprehensive software package, which contains the main Gaussian Process algorithms described in the according book (Rasmussen & Williams, 2006). Here the version GPML 3.0 has been used, which can be downloaded from http://gaussianprocess.org/gpml/code.

Curve Fitting Toolbox: A licensed toolbox by Mathworks for various regression and interpolation methods. Here, it was employed for NLR and LOWESS.

Furthermore, a software package called *mlhgp*, for the heteroscedastic Gaussian Process regression (HGP), has been used (Kersting *et al.*, 2007; Quadrianto *et al.*, 2009). This code directly builds on top of the GPML toolbox.

The interfaces for running experiments with the above toolboxes and for visualizing the results were implemented for the purpose of this diploma thesis. Also, a simple GP regression script (*sandbox*) which I wrote in the very beginning of this work.

The reader is encouraged to try the above toolboxes on his own, making use of the scripts which I added. For instruction, the functionalities of the most important implementations are summarized briefly in the following:

Functions	**Description**
`GPML/start_gpfit.m`	Interface for (multiple parallel) runs of Gaussian Process Regression using the GPML toolbox by C.E. Rasmussen and H. Nickisch (2010). Calls the function `GPML/gp_experiment.m` with seven parameters.
`GPML/gp_experiment.m`	Performs the actual GPML run, using several parameters such as the initial hyperparameters, the covariance function and an optimization flag.
`HGP/start_mlhgpfit.m`	Interface for running a Heteroscedastic Gaussian Process Regression (HGP) using the GPML toolbox by C.E. Rasmussen and H. Nickisch (2010) and the mlhgp code by K. Kersting (2007). Calls the function `HGP/mlhgp_experiment.m` with seven parameters. Written by Kristian Kersting (2007), extended by the author.
`HGP/mlhgp_experiment.m`	Performs the actual HGP run, using several parameters such as the initial hyperparameters, the number of latent noise variables and an optimization flag. Written by Kristian Kersting (2007), extended by the author.
`HGP/plot_mlhgp.m`	Plotting routine for a HGP experiment with 1D Input dimensionality. Makes use of the functions `HGP/binned_residuals.m` and `HGP/binned_residualsX.m`

Functions	**Description**
`HGP/plot_mlhgp_2D.m`	Plotting routine for a HGP experiment with 2D Input dimensionality. Makes use of the functions `HGP/binned_residuals.m` and `HGP/binned_residualsX.m`.
`sandbox/simple_gp.m`	Gaussian Process Regression script following solely the descriptions and equations in Bishop (2006) and Rasmussen & Williams (2006). Infers the posterior distribution, makes predictions and calculates the marginal likelihood, including a plotting routine. Makes use of a squared exponential covariance function (eq.2.28).

All the scripts documented here, as well as the open source toolboxes and an example dataset, are available on the enclosed CD. For further documentation, please refer to the `README.txt` file on the CD.

Acknowledgements

My supervisor Antje Maria Moffat deserves the greatest appreciation for constantly supporting and guiding the development of this diploma thesis. Her remarks, questions and challenges were of utmost constructiveness and value. I especially would like to thank her for the countless hours we spent in discussions and debates about the objectives, the structure and the scope of this thesis.
I acknowledge Prof. Schukat Talamazzini for charing the thesis. His hints, doubts, comments and helpful suggestions were fundamental for the progress of this work.
I am grateful to Markus Reichstein for his interest in the subject; his comments, ideas and questions were more than enlightening.
I thank Christian Beer for mentoring my first steps at a science institute.
Thanks to the entire Model Data Integration Group at the Max Planck Institute for Biogeochemistry in Jena. Altug, Bernhard, Gitta, Jannis, Maarten, Martin, Matthias, Miguel, Nuno, Thijs and Thomas for their helpful thoughts and questions about my topic, but also Angela, Myroslava and Uli for making our ivory tower a unique working environment.
Also, the Empirical Inference Group of Bernhard Schölkopf at the Max Planck Institute for Biological Cybernetics in Tübingen, for providing me with many insights into Machine Learning and Artificial Intelligence. Especially my supervisors Joris Mooij and Jonas Peters, but also Jakob, Jernej, Johannes, Kun and Max for sharing experiences, ideas and coffee breaks with me during the time of my internship.
I am thankful to Carl Edward Rasmussen for being incomparably enthusiastic and passionate about Gaussian Processes and Machine Learning. Besides, I am grateful to him, Hannes Nickisch and Kristian Kersting for developing excellent software packages which where essential for the progress of this thesis.
I would also like to thank Kerstin Hippler, for facilitating the visit at the flux towers in Hainich and Gebesee and for sharing useful information and "Freiland" know-how.
The computer service staff at the MPI in Jena for the friendly cooperation and for providing computational power on a professional level. At most Peer-Joachim Koch, who was always willing to help in problems regarding the HPC or programming in general.
For proof-reading and giving valuable comments on language and grammar I thank Myroslava Khomik and Gunnar Menzer.
My family and my friends deserve special thanks for their moral support and encouragement. Finally, I thank my parents for guiding my ways through education, but still providing me all the freedom to choose a study path of my own interest.

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Jena, den 05.08.2011 .....

(Olaf Menzer)