Statistical Approaches to Leak Detection for Geological Sequestration

by

Arman S. Haidari

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Chemical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

April 2011
[June 2011]

Author .....
Department of Chemical Engineering
April 8, 2011

Certified by .....
Gregory J. McRae
Hoyt C. Hottel Professor of Chemical Engineering
Thesis Supervisor

Accepted by .....
William M. Deen
Chairman, Committee for Graduate Students

Statistical Approaches to Leak Detection for Geological Sequestration

by

Arman S. Haidari

Submitted to the Department of Chemical Engineering
on April 8, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Chemical Engineering

Abstract

Geological sequestration has been proposed as a way to remove CO_2 from the atmosphere by injecting it into deep saline aquifers. Detecting leaks to the atmosphere will be important for ensuring safety and effectiveness of storage. However, a standard set of tools for monitoring does not yet exist.

The basic problem for leak detection - and eventually for the inverse problem of determining where and how big a leak is given measurements - is to detect shifts in the mean of atmospheric CO_2 data. Because the data are uncertain, statistical approaches are necessary. The traditional way to detect a shift would be to apply a hypothesis test, such as Z- or t-tests, directly to the data. These methods implicitly assume the data are Gaussian and independent. Analysis of atmospheric CO_2 data suggests these assumptions are often poor. The data are characterized by a high degree of variability, are non-Gaussian, and exhibit obvious systematic trends. Simple Z- or t-tests will lead to higher false positive rates than desired by the operator. Therefore Bayesian methods and methods for handling autocorrelation will be needed to control false positives.

A model-based framework for shift detection is introduced that is capable of coping with non-Gaussian data and autocorrelation. Given baseline data, the framework estimates parameters and chooses the best model. When new data arrive, they are compared to forecasts of the baseline model and testing is performed to determine if a shift is present. The key questions are, how to estimate parameters? Which model to use for detrending? And how to test for shifts?

The framework is applied to atmospheric CO_2 data from three existing monitoring sites: Mauna Loa Observatory in Hawaii, Harvard Forest in central Massachusetts, and a site from the Salt Lake CO_2 Network in Utah. These sites have been chosen to represent a spectrum of possible monitoring scenarios. The data exhibit obvious trends, including interannual growth and seasonal cycles. Several physical models are proposed for capturing interannual and seasonal trends in atmospheric CO_2 data. The simplest model correlates increases in atmospheric CO_2 with global annual emissions of CO_2 from fossil fuel combustion. Solar radiation and leaf area index models are proposed as alternative ways to explain seasonality in the data.

Quantitative normality tests reject normality of the CO_2 data and the seasonal models proposed are nonlinear. A simple reaction kinetics example demonstrates that nonlinearity in the detrending model can lead to non-Gaussian posterior distributions. Therefore Bayesian methods estimation will be necessary. Here, nonlinear least squares is used to reduce computational effort.

A Bayesian method of model selection called the deviance information criterion (DIC) is introduced as a way to avoid overfitting. DIC is used to choose between the proposed models and it is determined that a model using a straight line to represent emissions driven growth, the solar radiation model and a 6-month harmonic term does the best job of explaining the data. Improving the model is shown to have two important consequences: reduced variability in the residuals and reduced autocorrelation.

Variability in the residuals translates into uncertainty in CO_2 forecasts. Thus by reducing the spread of the residuals, improving the model increases the signal to noise ratio and improves the ability to detect shifts.

A least squares example using CO_2 data from Mauna Loa is used to illustrate the effect of autocorrelation due to systematic seasonal variability on the ability to detect. The issue is that ordinary least squares tends to underestimate uncertainty when data are serially correlated, implying high false positive rates. Improving the model reduces autocorrelation in the residuals by eliminating systematic trends. Because the data exhibit gaps, Lomb periodograms are used to test the residuals for systematic signals. The model chosen by DIC removes all of the growing and seasonal trends originally present at the 5% level of significance. Thus improving the model is a way to reduce autocorrelation effects on false positives.

A key issue for future monitoring sites will be demonstrating the ability to detect shifts in the absence of leaks. The urban weekend weekday effect on atmospheric CO_2 is introduced to illustrate how this might happen. A seasonal detrending model is used to remove systematic trends in data at Mauna Loa, Harvard Forest and Salt Lake. Residuals indicate the presence of positive shifts at the latter sites, as expected, with the magnitude of the shift being larger at the urban site than the rural one ($\sim$ 8 ppm versus $\sim$ 1 ppm). Normality tests indicate the residuals are non-Gaussian, so a Bayesian method based on Bayes factors is proposed for determining the amount of data needed to detect shifts in non-Gaussian data. The method is demonstrated on the Harvard and Salt Lake CO_2 data. Results obtained are sensitive to the form of the error distribution. Empirical distributions should be used to avoid false positives. The weekend weekday shift in CO_2 is detectable in 48-120 samples at the urban site. More samples are required at the rural one. Finally, back-of-the-envelope calculations suggest the weekend weekday shift in emissions detected in Salt Lake is $\sim O(0.01)$ $MtCO_2km^{-2}yr^{-1}$. This is the equivalent of 1% of 1 $MtCO_2$ stored belowground leaking over an area of 1 km^2.

The framework developed in this thesis can be used to detect shifts in atmospheric CO_2 (or other types of) data after data is already available. Further research is needed to address questions about what data to collect. For example, what sensors should be used, where should they be located, and how frequently should they be sampled? Optimal monitoring network design at a given location will require balancing the need to gather more information (for example, by adding sensors) against operational constraints including cost, safety, and regulatory requirements.

Thesis Supervisor: Gregory J. McRae
Title: Hoyt C. Hottel Professor of Chemical Engineering

Acknowledgments

This dissertation would not have been possible without the support and guidance from several important individuals.

First and foremost I offer my sincerest gratitude to my thesis advisor, Professor Gregory McRae, who has shared with me his extensive knowledge and who has pushed me to hold myself to the highest standards. Without him, this work would not have been possible.

I am grateful to my committee members for their patience and guidance: Professor Bill Green for challenging me to look at my work from a new perspective, Professor Howard Herzog for so generously donating his time and expertise on short notice, and Professors Ronald Prinn and Ruben Juanes for their ideas and insights.

I would like to thank my lab members for their cooperation and support throughout this process: Sara Passone, Chuang-Chung Lee, Alex Lewis, Mihai Anton, Ingrid Berkelmans, Carolyn Seto, Bo Gong, Ken Hu, and Kunle Adeyemo.

I am grateful for the help of department administrators Suzanne Maguire and Katie Lewis, the computer support of Jean Belbin and Jim Hardsog who have saved me countless time and effort, and especially Joan Chisholm for her unconditional support and advice over the years.

Special thanks is owed to Steve Wofsy and Bill Munger at Harvard University and Jim Ehleringer and Steve Rowley of the University of Utah for their input and for sharing data they have worked hard to generate.

Finally, I would like to acknowledge the MIT Carbon Sequestration Initiative for their financial support.

It takes a village to raise a child. I dedicate this thesis to my village: to my parents, Rose and Mehdi; to my sister Eman for always being there for me even when I couldn't be for her; to Mina for her love; to Jaleh for her selfless support; to my new family, Jasmine, Nima, Ehsan, and Gerry; and for the countless friends and supporters I have and have had over the years. I thank you all.

Contents

List of Figures

List of Tables

Notation

Table 1: Conventions for notation

Symbol(s)	Description
	Scalars, vectors and matrices
a	A scalar.
$\mathbf{a}$	A column vector.
A	A set, as in $A = a_1, a_2, ..., a_3$.
$\mathbf{A}$	A matrix.
$a(\omega)$	Random variable
$\hat{a}$	Point estimate of $a(\omega)$; for example, a least squares estimate or the mode from Bayesian estimation.
$\boldsymbol{A}^T$	Transpose of $\boldsymbol{A}$.
	Observations and model predictions
t	Time.
y, y_{obs}	Observation.
x	State variable.
θ,β	Model parameter, linear regression coefficient.
t_s	Time of shift in mean CO_2 associated with leak.
Δy	Magnitude of shift in mean CO_2 associated with leak.
ρ_i	Autoregressive model coefficient
ϕ_i	Moving average model coefficient
$y(\theta, t\|\mathcal{M})$,$\mathcal{M}(\theta, t)$	Model approximation to nature, for predicting CO_2.
	Errors
ξ	Random error, also known as a random innovation.
ε_{obs}	Observation error due to instrumentation and averaging.
$\varepsilon_{\mathcal{M}}$	Error from approximating nature with model $\mathcal{M}$
$\varepsilon_\theta, \varepsilon_{\Delta y}, \varepsilon_{t_s}$	Errors due to uncertainty.
$\varepsilon = \varepsilon_{\mathcal{M}} + \varepsilon_{obs}$	Model error.

Continued on Next Page...

Table 1 – Continued

Symbol(s)	Description
$\varepsilon = \varepsilon + \varepsilon_\theta$	Forecast error.
$e, \Delta y(t)$	Residual. An outcome of ε. $\Delta y(t)$, not to be confused with Δy which is a shift parameter, is sometimes used in place of e.
	Estimates and other important terms
$\boldsymbol{J}$	Jacobian matrix.
∇	Gradient vector.
$\boldsymbol{V}_\theta, \boldsymbol{V}_\beta$	Variance-covariance matrix of θ or β.
μ	Population mean.
σ^2	Population variance.
ρ	Population correlation coefficient.
$\hat{\mu}$	Sample mean.
$\hat{\sigma}^2, s^2$	Sample variance.
$\hat{\rho}$	Sample correlation coefficient.
	Shift window and averaging
T	Time period, usually the length of the window of time considered to detect a shift.
N	Number of observations considered for detection.
τ	Averaging time, also used as a lag coefficient.
	Probability
$\pi(\theta)$	Prior density, sometimes used to represent a generic density function for θ.
$\pi(\theta\|y_{obs})$	Posterior density.
$f(y\|\theta)$ or $\ell(\theta)$	Likelihood function.
$L(\theta)$	Natural logarithm of the likelihood function.
$f(y), f(y\|\mathcal{M})$	Predictive density for model $\mathcal{M}$, sometimes called the marginal of the observations because it is the result of marginalizing $f(y, \theta)$ over the unknowns.
$p(\cdot)$	Probability mass function.
$\mathrm{P}(\cdot), \Pr(\cdot)$	Cumulative probability
$\mathrm{E}[\cdot]$	Expected value, mean
$\mathrm{Var}\{\cdot\}$	Variance
$\mathcal{N}, \mathscr{N}$	Normal distribution

Acronyms

Table 2: List of acronyms

Acronym	Term
CCS	Carbon capture and storage
RSS	Sum of squared residuals
AIC	Akaike's information criterion
BIC	Bayesian information criterion, also called Schwarz's information criterion
DIC	Deviance information criterion
K-L	Kullback-Leibler
K-S	Kolmogorov-Smirnov
B-G	Breusch-Godfrey
AR	Autoregressive error model
MA	Moving average error model
ARMA	Autoregressive moving average error model
ppm	Parts per million, sometimes referred to as ppmv for parts per million by volume; a mole fraction
SNR	Signal to noise ratio
FPE	Final prediction error
R^2	Coefficient of determination
OLS	Ordinary least squares
(E)GLS	(Estimated) generalized least squares
NOAA	National Oceanic and Atmospheric Administration
SIO	Scripps Institution of Oceanography
MAP	Maximum a posteriori
WMO	World Meteorological Organization
PFC	Perfluorocarbon (a class of tracer molecules)

Chapter 1

Introduction

1.1 Motivation

1.1.1 Carbon dioxide and climate change

There is broad agreement within the scientific community that anthropogenic emissions, especially the release of CO_2 from fossil fuel combustion, are responsible for recent shifts in climate and that if left unchecked, these emissions could have serious consequences [1] for the future. Continuous records of atmospheric CO_2 were initiated by C.D. Keeling at Mauna Loa in 1958 [2]. Before 1950, the CO_2 record is based primarily on indirect measurements of air bubbles trapped in polar ice. The ice core records indicate that historic CO_2 mixing ratios ranged between 180 and 300 ppm [3]. Between 1750 and 2005, the global average CO_2 mixing ratio increased from somewhere between 275 and 285 ppm to 379 ppm, with half the change coming in the last ~30 years [4]. The global average CO_2 mixing ratio was approximately 386 ppm in 2009 and has been increasing at a rate of 0.66-2.92 ppm yr^{-1} since 1980 [5]. Measurements of atmospheric $^{13}C/^{12}C$ isotopic ratios (see [6, 7, 8, 9, 4]) and atmospheric O_2 (see [10, 11, 12, 4]) have made possible the partitioning of CO_2 fluxes to the atmosphere into separate fossil fuel, ocean and land contributions. It is estimated that more than 75% of the shift in atmospheric CO_2 since pre-industrial times is due to fossil fuel combustion and cement manufacture [13]. This trend is unlikely to change, with international emissions projected to increase from 29.7 $GtCO_2$ (in 2007) to 42.4 $GtCO_2$ by 2035. The same forecasts [14] predict a growing fraction of these emissions will be due to coal, with coal representing 46% of fossil fuel emissions in 2035. The bulk of this increase is projected to come from non-OECD (Organisation for Economic Co-operation and Development) countries, especially (but not solely) China, with coal consumption rates in 2035 forecasted at about three times what they were in 1990. OECD countries are expected to increase their rates of natural gas consumption, with coal and petroleum remaining roughly constant.

1.1.2 Mitigating CO_2 emissions

Many approaches have been proposed for reducing anthropogenic CO_2 emissions to the atmosphere. These fall into a few main categories: (1) increasing energy efficiency of existing appliances, automobiles, buildings, etc., (2) changing behavioral patterns to conserve energy, (3) using carbon free technologies like nuclear or switching to reduced carbon energy sources like natural gas, and (4) capturing and storing CO_2 before it can be emitted to the atmosphere.

It is likely that a portfolio of technologies will be required to make the needed reductions in CO_2 emissions (see, for example, [15]). Pacala [16] reviews several large scale options for reducing CO_2 emissions, each of which could reduce CO_2 emissions by 1 $GtCyr^{-1}$ (=3.67 $GtCO_2yr^{-1}$). Among these, three fall under the umbrella of carbon capture and storage (CCS), and one of these three entails capturing and storing CO_2 from a coal or natural gas power plant. A single 500 MW coal fired power plant releases ~3 $MtCO_2yr^{-1}$ to the atmosphere [17]. One of Pacala's wedges would therefore require capturing and storing CO_2 from over a thousand, 500 MW coal fired power plants. The U.S. alone has more than 500 such plants. CCS from coal fired and natural gas fired power plants is a particularly attractive option in view of the fact that coal and natural gas based CO_2 emissions are likely to continue to increase for decades to come.

1.1.3 Carbon capture and storage

Carbon capture and storage entails capturing carbon dioxide from large scale point sources and then isolating it from the atmosphere by storing it. Four basic approaches have been proposed for storing CO_2: mineral carbonation, industrial use, ocean storage, and geological sequestration [18]. Mineral carbonation involves chemically processing CO_2 to convert it into inorganic carbonates. This technology is energy intensive, costly and would have significant environmental impacts of its own. Industrial uses of CO_2 involve short term storage in intermediate forms like carbonation in drinks and urea. Total industrial usage of CO_2 only consumes on the order of 100 $MtCO_2yr^{-1}$ [18], not nearly the demand that would be required for a large scale option [16]. Ocean storage could take one of three forms: column release, sea floor release, or dissolution of mineral carbonates. Experience with these forms of storage is extremely limited and more research is needed. The best CO_2 mitigation option is geological sequestration.

1.1.4 Geological sequestration and leak detection

Geological sequestration involves injecting CO_2 belowground into one of several types of formations: a depleted hydrocarbon field, a saline aquifer, a salt cavern, or an active hydrocarbon field. Saline aquifers are the best option for mitigation, with storage capacities estimated in the thousands of Gigatonnes. While underground CO_2 storage is not new - enhanced recovery, for example, has been performed for decades now - scale is a concern. Today, just a handful of sites inject $O(1)$ $MtCO_2yr^{-1}$.

Performing geological sequestration at the Gigatonne scale will require the equivalent of thousands of such sites. Such dramatic scaling up means engaging in carbon capture and storage well beyond the current operating range. There is concern about what might happen if CO_2 leaks into groundwater or the atmosphere. Several leakage pathways exist, including already existing fractures, faults and wellbores and also new pathways due to dissolution of rock by acidic CO_2 rich fluids and mechanical stresses associated with injection [19, 20, 21]. Clearly there is a need for monitoring to detect leaks. Today, a standard set of protocol for monitoring does not exist. This thesis is geared toward understanding how detection should take place at future geological sequestration monitoring sites.

1.2 Thesis statement

Geological sequestration is being proposed as a way to remove CO_2 from the atmosphere by injecting carbon dioxide into deep saline aquifers. A key challenge is to ensure that the CO_2 remains underground and does not leak into the atmosphere. Monitoring will be needed to detect leakage to ensure safety and effectiveness of storage. However, a standard procedure for monitoring to detect leaks does not exist.

At its most basic, the leak detection problem is to detect upward shifts in atmospheric CO_2 concentration data. A few obvious questions arise:

- If there were a sudden (or even a slow) leak, how long would it take to detect a statistically significant change?
- From a policy perspective, what are the risks associated with a false positive signal detection?
- If a monitoring system were to be put in place, how could it be tested if there were no leaks to the atmosphere?

This thesis is directed at developing the statistical tools, analysis methodologies and network design decisions needed to address each of the above issues.

1.3 Chapter overview

The discussion in the remainder of this thesis proceeds as follows:

- **Chapter 2.** The leak detection problem is introduced in the context of atmospheric CO_2 concentration data from the Mauna Loa Observatory. The data are characterized by systematic seasonal cycles and growing trends linked to biology and fossil fuel emissions. Superimposed on these systematic trends there are deviations primarily due to averaging and instrumental calibration errors. The challenge is to detect shifts in such data associated with leaks.

  Because the data are uncertain, statistical approaches are appropriate. Traditional hypothesis tests, such as Z- and t-tests, can be applied. False positive

rates can be reduced by increasing the signal to noise ratio or by providing more time for data analysis. Larger signal to noise ratios can be achieved by improving the accuracy and precision of models and/or sensors. A statistical framework is introduced for combining models and data to detect shifts. Statistical approaches to shift detection are introduced. Two factors affect false positive rates: normality and autocorrelation.

CO_2 data are often not normally distributed. Exploratory statistics and quantitative normality tests are used to show CO_2 data from one site is non-Gaussian. Z- and t-tests implicitly assume the data are Gaussian. Blindly using such tests when data are not normally distributed will lead to false positive rates that are different in practice from the level chosen by the decision maker. Methods of parameter estimation are introduced and a chemical reaction example is used to show that nonlinear models can lead to non-Gaussian posterior distributions. The consequence is that non-Bayesian estimation methods will tend to misrepresent uncertainty in future atmospheric CO_2, leading to loss of control over false alarm rates just as occurs when Z-tests are applied non-Gaussan data. Bayesian methods are needed to cope with non-Gaussian uncertainties.

Autocorrelation is an issue because standard estimation methods such as least squares, which form the basis for the standard errors most often reported with sample means, tend to underestimate uncertainty when data are serially correlated. The implication is that methods such as generalized least squares should be used for parameter estimation when modeling errors are correlated; otherwise false alarm rates will be higher than they should be. Model building, estimation and selection and the effects of modeling on autocorrelation are discussed later in the context of atmospheric CO_2 from existing monitoring sites.

- **Chapter 3.** Two examples are used to illustrate the effects of non-normality and autocorrelation on false positives.

  The first example involves random trajectories having different distributions but the same means and variances. The Gaussian is used to illustrate the impact of waiting longer ($N\Delta t$) and the effect of forward (parameteric) uncertainty on the ability to detect are both illustrated. There is a tradeoff between the time allotted for detection and the certainty with which conclusions about shifts can be made. Issues associated with data being non-Gaussian are illustrated. The Central Limit Theorem may not apply; confidence intervals assuming Gaussian data will lead to false positives. The non-Gaussian data motivate the need for Bayesian methods.

  The second example uses CO_2 data from Mauna Loa to demonstrate the problem with ordinary least squares when data are autocorrelated. Later it is shown that modeling can remove the systematic trends contributing to correlation. In general, modeling may not resolve all correlation in the data and methods akin to generalized least squares will be needed to avoid false positives.

- **Chapter 4.** Three data sources are chosen representing a cross section of

the types of settings in which monitoring would take place in the future. The first source of data, Mauna Loa Observatory, is used to represent an ideal site, where diurnal variability is minimal. The other two sites - Harvard Forest and the University of Utah in Salt Lake City, Utah - represent locations where local biology and urban emissions lead to greater variability in the data. Key findings from past studies of atmospheric CO_2 at Mauna Loa, Harvard Forest and Salt Lake are summarized and combined with exploratory statistics to give implications for modeling.

- **Chapter 5.** Several physically inspired, correlative models are built for predicting interannual and seasonal changes in atmospheric CO_2 using the findings from Chapter 4. Improved models decrease the magnitudes of the residuals by explaining more of the systematic variability in the data. While for shift detection decreasing the variance of the residuals is desirable because it increases the signal to noise ratio, there is nothing to stop an operator from adding terms until the model perfectly fits the data. The deviance information criterion (DIC) - a Bayesian method of model selection which balances goodness-of-fit with parsimony - is introduced as a way to avoid overfitting. Applying the criterion to validation data at Mauna Loa, Harvard Forest and Salt Lake City suggests using a simpler extended solar radiation model of seasonality rather than a more complex model based on leaf dynamics. The final model predicts seasonal changes well at all three sites considered.

- **Chapter 6.** Forecast uncertainties are quantified for each site. It is shown that the bullk of forecast uncertainty at sites influenced by local sources and sinks arises from diurnal variability when seasonal models are used.

- **Chapter 7.** A key issue is demonstrating the ability to detect a leak at new monitoring sites where leaks have not occurred. The urban weekend weekday shift detection problem is used to illustrate how this might happen in practice. The data are non-Gaussian. Therefore a Bayesian method of detection is introduced. The method is applied to data from Harvard Forest and Salt Lake City. The results indicate that less time would be required to detect the shift present at Salt Lake. This makes sense, as the Harvard site is far from metropolitan areas. A back-of-the-envelope calculation suggests the shift detected at Salt Lake is on the order of 10,000 tonnes CO_2 km^{-2} yr^{-1}. This is the equivalent of 1% of 1 $MtCO_2$ stored belowground leaking over an area of 1 km^2.

- **Chapter 9.** Conclusions relevant to leak detection for geological sequestration are discussed.

- **Chapter 10.** Future research directions are given.

Chapter 2

Statistical approaches to leak detection

This chapter introduces the leak detection problem mathematically. Different statistical frameworks for detection are briefly described, and issues affecting detection are highlighted. Finally, lessons applicable to the leak detection problem are summarized.

2.1 Which variable to monitor?

For leaks to be detected, the variable(s) being monitored should be sensitive to leakage. Examples of near-surface variables that could be used include soil gas flux, perfluorocarbon (PFC) tracer concentrations, isotopic abundances of $\delta^{13}C$ and $\delta^{14}C$ in soil gas, pH of groundwater, and atmospheric CO_2 concentration [19, 22]. It is unclear what variables should be monitored. A sensor might provide lots of information about leakage, but if this information is expensive to obtain then cheaper less informative sensors could be preferable. Optimal monitoring designs addressing such tradeoffs are an important topic for future research. This thesis focuses on how to analyze a given data stream for changes related to leakage.

Figure 2-1 shows atmospheric CO_2 concentration data from Mauna Loa Observatory. Several characteristics are apparent in the data. First, there is a small growing trend of magnitude $\sim$1 ppm per year. Superimposed on this growing trend is a seasonal cycle whose peak to peak amplitude is roughly 5 ppm. On top of these systematic trends there are random deviations ranging between ± 2 ppm due primarily to averaging and instrumental calibration errors. The leak detection problem is to detect a shift in noisy data like those in the figure.

While atmospheric CO_2 concentration is just one of many examples of variables that could be targeted for monitoring to detect leaks, it is a useful variable to consider because of the high degree of natural variability present. This variability, which has both random and systematic components and is characteristic of both fluxes and concentrations of CO_2 near the Earth's surface, necessitates careful consideration of trends and uncertainties during decision making. The remainder of this thesis therefore uses atmospheric CO_2 concentration data as a vehicle for describing how to

detect leaks in the presence of uncertainty. The ideas and methods introduced in the context of atmospheric CO_2 are equally applicable to situations where fluxes or other variables are targeted for monitoring.

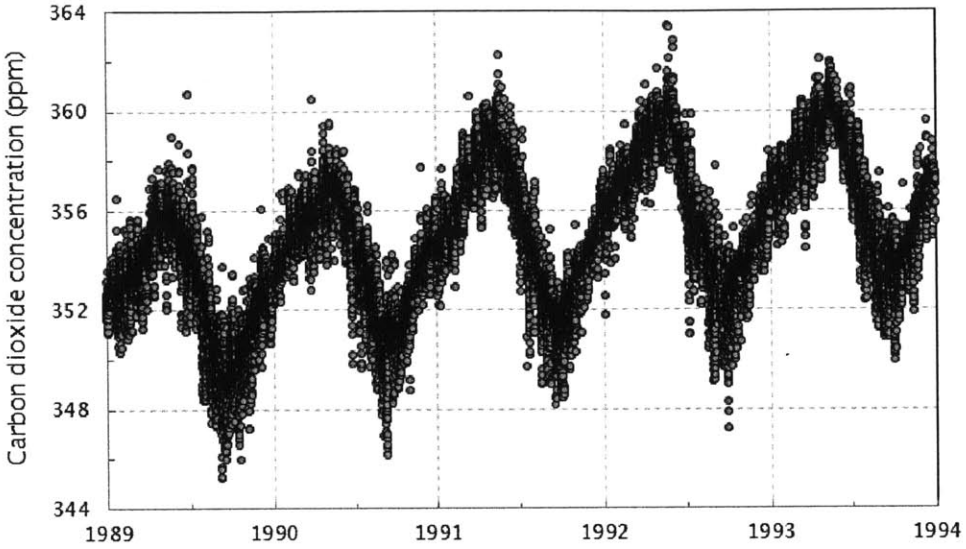

Figure 2-1: Atmospheric CO_2 data from Mauna Loa Observatory, Hawaii (20°N, 156°W) where CO_2 concentration is in parts per million in the mole fraction (ppm).

2.2 Fluxes and concentrations

Shifts in atmospheric CO_2 concentration provide information about fluxes at the surface and as such can be used to detect leaks from belowground. Figure 2-2 illustrates a uniform leak flux of magnitude E moles per $m^{-2}s{-}1$ into a well mixed box into which air is entering on the left at background concentration CO_2^b moles m^{-3} and leaving on the other side at concentration CO_2 moles m^{-3}. Because CO_2 is leaking into the box, the concentration leaving will be higher than that entering. At steady state, the perturbation in carbon dioxide concentration across the box is given by

$$\begin{aligned} \Delta CO_2 &= CO_2 - CO_2^b \\ &= \frac{LE}{uz_i} \end{aligned} \tag{2.1}$$

where L is related to the spatial scale of the leak and z_i is the mixing height dictated by boundary layer dynamics. Equation (2.1) states that the size of the shift in atmospheric CO_2 concentration across the leak is directly proportional to the magnitude of the leak flux. If sensors were placed upwind and downwind of the leak, then the difference in their signals will reflect leakage.

The atmosphere is a dynamic medium. Figure 2-3 illustrates what happens when wind direction changes. Initially, the wind carries leaking CO_2 to a sensor measuring

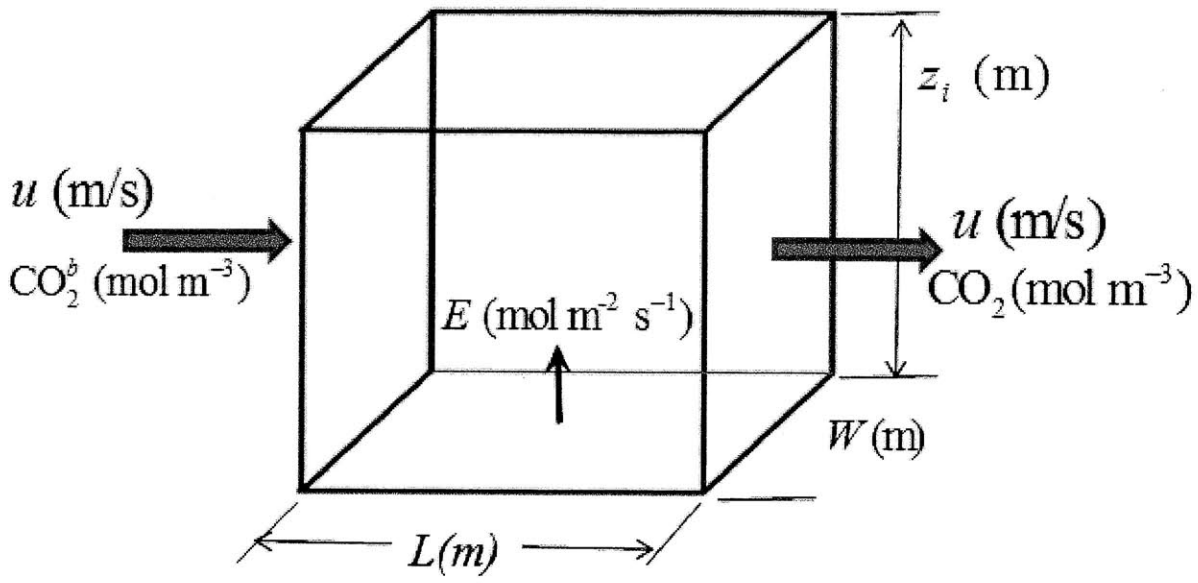

Figure 2-2: A simple well mixed box model.

atmospheric CO_2 concentration, marked in blue. At some later time the wind direction shifts, carrying the plume away from the sensor. The trajectory followed by the sensor data shifts downward. Placing additional sensors in the field, as shown in the figure on the bottom left, can ensure detection regardless of wind direction. As wind direction changes, the data obtained from sensors in the field will shift up and down. Whether a single sensor or multiple sensors are used, the ability to detect shifts in atmospheric CO_2 concentration is critical for detecting leaks and inferring how big they are and where they are located.

2.3 Detecting mean shifts in atmospheric CO_2

Figure 2-4 illustrates the effect of an artifical +20 ppm step change after 00:00:00 July 3, 1991, in the Mauna Loa CO_2 data from Figure 2-1. The exact time profile of the CO_2 perturbation will depend on the spatial and temporal profile of the leak and on atmospheric mixing between the leak and the sensor. In practice, gradual shifts evolving more smoothly over time than step changes can be expected. Detecting gradual shifts will require more time.

2.4 Conceptual methods for detecting shifts in CO_2

The leak detection problem is a mean shift detection problem. Mean shifts arise under many different names in the literature [23], including observational inhomogeneities, change points, and regime shifts. Homogenization is the practice of detecting and

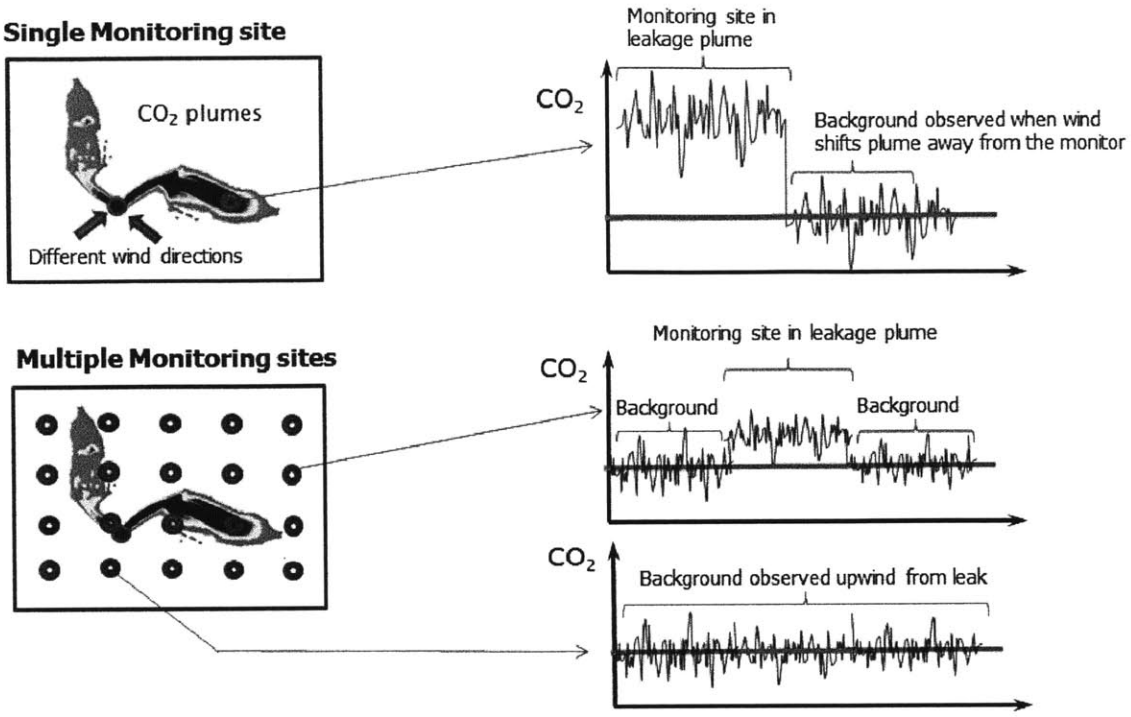

Figure 2-3: Challenges in surface monitoring.

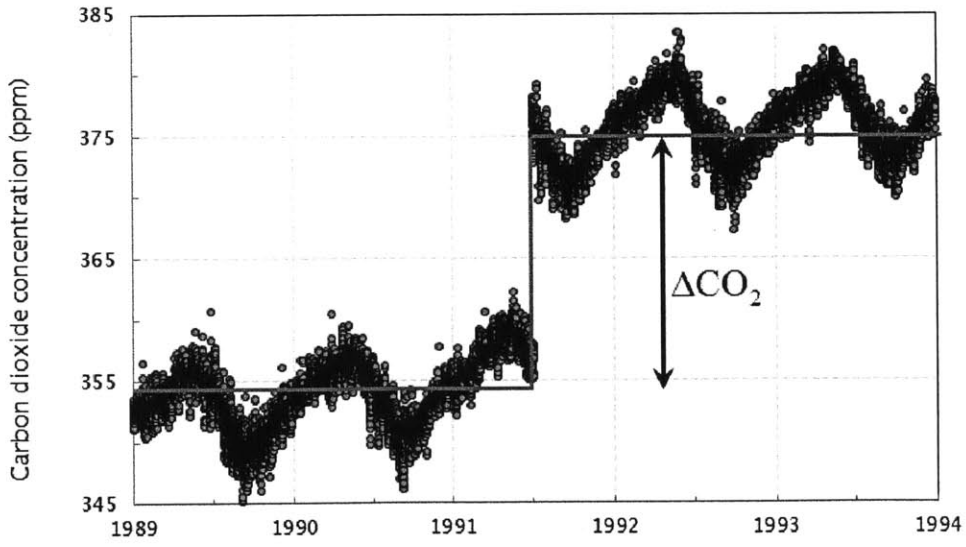

Figure 2-4: A hypothetical +20 ppm shift in atmospheric CO_2 data from Mauna Loa Observatory.

then correcting for artificial shifts in time series and is commonly used in climate science. Regime shift detection methods arise in finance, economics and ecology where there is interest in issues such as the onset of economic recession and changes in population dynamics. The changes being sought out can be mean shifts, shifts in variance, changes in harmonic content, or structural changes in the system governing the observations [24].

Figure 2-5 illustrates three possible ways to think about detecting shifts in mean CO_2. The idea behind all three approaches is to slide a window across the data and at each point along the series analyze the data in the window. Averaging leads to moving averages like the one shown in the figure on the top right, where the moving average shifts upward in the vicinity of the shift time. Computing variances, the shift will result in a spike in the series of variances obtained. Finally, if histograms or density functions are generated then the upward shift in the data will result in a leftward shift in the distribution after the onset of a shift associated with leakage. This last approach - detecting shifts in distributions - is one that will be revisited later in the thesis.

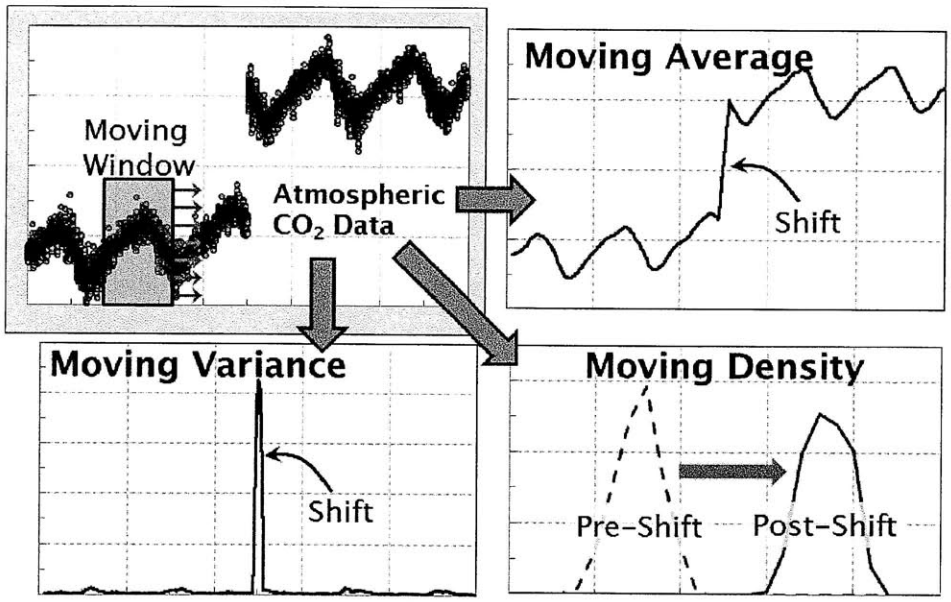

Figure 2-5: Ideas for mean shift detection.

Figure 2-6 shows how the sample mean evolves as the size of the averaging window increases. Oscillations present due to natural variability are damped out and the mean value converges to its long term value as more data is analyzed. The result is less biased estimates of the mean and therefore reduced risk of false positives.

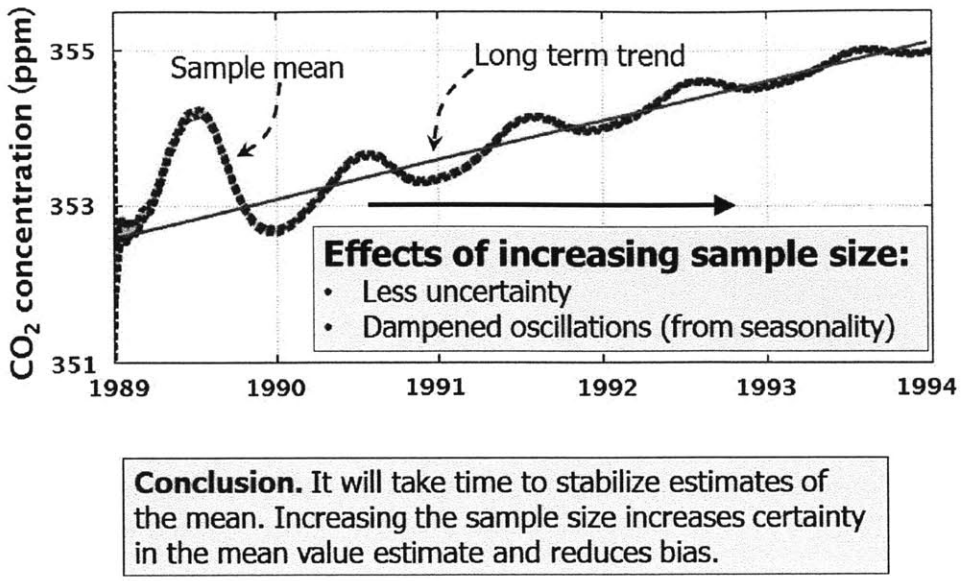

Figure 2-6: Convergence of the sample mean as window size increases.

2.5 Uncertainty in the data

The data used in this thesis are time averages of atmospheric CO_2 concentration. There are two important points to be made in regard to the data. First, they are uncertain in the sense that the true value being measured is unknown and will always error from the measured value by some amount. Second, there are two sources of error: averaging and instrumental calibration.

Most of the data used in this thesis are hourly averages. Figure 2-7 illustrates the procedure leading to the average from a hypothetical hour. Typically several data are recorded within the hour, which are referred to in the figure as 5-minute averages. The hourly average is then obtained as the mean of these subhourly data, as in

$$y(t) = \frac{1}{n_{avg}} \sum_{i=1}^{n_{avg}} y_{5min}(t_i) \tag{2.2}$$

Each 5-minute datum errors from the true value because of calibration error. [1] The calibration procedure requires conversion of raw voltages into atmospheric concentrations. This is done through a calibration model, which is never perfect. As a result, the 5-minute data will tend to error from the true value by fixed amounts. Because these amounts are unobservable, they can be treated as random errors. The random errors in the subhourly data persist in the hourly average. However, calibration is not the only source of error. Because an hourly average cannot capture subhourly trends in nature, hourly values will tend to differ from their true counterparts even

[1] Drift can be ignored because the data used in this thesis come from instruments that are calibrated frequently (on an hourly basis).

in the absence of calibration error. Again, these differences are unobservable and can be treated as random errors. Therefore, hourly data can be modeled as

$$y(t) = y_{nature}(t) + \varepsilon_o(t) \tag{2.3}$$

where $y_{nature}(t)$ is the true CO_2 value and

$$\varepsilon_o(t) = \varepsilon_{instrument}(t) + \varepsilon_{averaging}(t) \tag{2.4}$$

is the error associated with a given hourly datum. The question is, what is the scale of the hourly error?

Scales of errors are most often reported in terms of standard deviations. For the sites considered here, two different types of standard deviations are reported. One type is the standard deviation of the subhourly samples with regard to the hourly mean value. In this case, the standard deviation is the square root of the sample variance determined using

$$s^2(t) = \frac{1}{n_{avg} - 1} \sum_{i=1}^{n_{avg}} [y_{5min}(t_i) - y(t)]^2 \tag{2.5}$$

The other type is the standard deviation associated with the hourly value $y(t)$. This is determined as the square root of

$$s_y^2(t) = \frac{s^2(t)}{n_{avg}} \tag{2.6}$$

Because the instrumental and averaging errors are random,

$$\begin{aligned} \text{Var}\{y(t)\} &= \text{Var}\{\epsilon_{instrument}(t)\} + \text{Var}\{\epsilon_{averaging}(t)\} \\ &= \sigma^2_{instrument} + \sigma^2_{averaging} \end{aligned} \tag{2.7}$$

where $\sigma^2_{instrument}$ and $\sigma^2_{averaging}$ denote the variances of the errors due to calibration and averaging. The uncertainty in the data will tend to grow as the instrument becomes less precise and as natural variability during an hourly interval increases. Natural variability will depend on location. The instruments used to generate the data used here are typically very precise, with standard deviations less than 1 ppm. Because the data are uncertain, statistical approaches to detection are useful. The next section describes popular approach based on hypothesis testing.

2.6 Detecting shifts with Z- and t-tests

Z- and t-tests are the most commonly used methods for comparing the means of two populations. The tests (see Table 2.1) operate in a nearly identical fashion. Data from one of the populations of interest is used to compute a mean value. This is compared to the mean value of the other population. If the difference is extreme

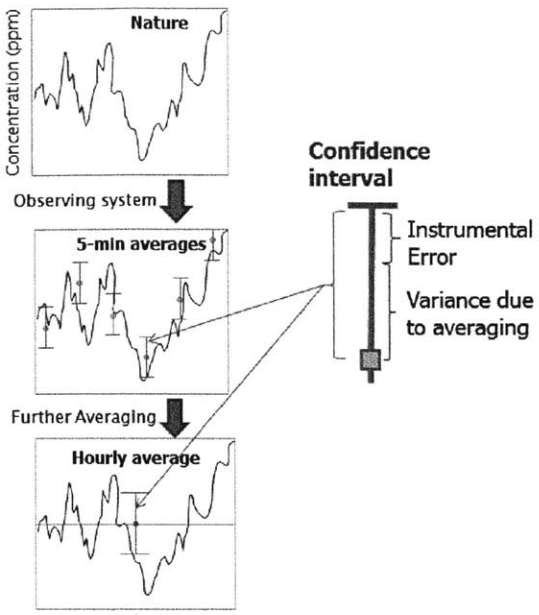

Figure 2-7: Averaging and instrumental errors in CO_2.

given the natural variability in the data, then a shift is concluded present.

To demonstrate the factors affecting the ability to detect shifts, consider a simple example. Assume CO_2 data up to some point in time are independent and identically distributed according to $N(a_0, a_1^2)$, where a_0 and a_1^2 are the known mean and variance, respectively. A leak signal may or may not already be present. If the variance of the data is not affected by the shift induced by leakage, then a Z-test can be used to test for shifts given new data $\boldsymbol{y} = [y_1, ..., y_N]'$. The Z-test starts by computing the statistic

$$\mathcal{T} = \frac{\bar{\boldsymbol{y}} - a_0}{a_1/\sqrt{N}} \tag{2.8}$$

where $\bar{\boldsymbol{y}}$ is the sample mean of the new data. If the data are indeed normally distributed[2], then the test statistic $\mathcal{T}$ will have a standard normal distribution in the absence of a shift. A shift is declared whenever the value of the test statistic is more extreme than a critical value z. The value of z is dictated by the **level of significance** (α) required by the user. Because the data are uncertain, there is a finite probability of concluding a shift has occurred when none is present. This event is called a false positive or false alarm. False positives are obviously events any operator would like to avoid, to minimize effort spent diagnosing erroneous shifts. The level of significance α is the false positive rate the operator is willing to live with. For example, if $\alpha = 0.05$, then false positives will occur on average 5% of the time shifts are detected.

[2]If the individual data are Gaussian, then this assumption is satisfied. If not, then by the Central Limit theorem it will be for a sufficiently large number of observations. The implication of the Central Limit theorem is discussed shortly.

Suppose the operator specifies the shift size to be detected at the α level of significance is $\Delta y = k \cdot a_1$. k is the signal to noise ratio (SNR) of the shift. The probability of false alarm is related to the SNR according to

$$\alpha = 1 - \text{erf}\left(\frac{k}{\sqrt{2N}}\right) \tag{2.9}$$

where erf represents the error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp -t^2 dt. \tag{2.10}$$

The conclusion is that the level of significance (and the false positive rate) decreases (1) as the number of data considered increases and (2) as the signal to noise ratio increases. Therefore strategies for reducing false positive rates include:

1. Increasing the signal to noise ratio (SNR) by
   - removing systematic sources of variability in the data and/or
   - improving the accuracy and precision of measurements; or
2. Increasing the number of observations (N) by
   - waiting longer ($N\Delta t$) and/or
   - adding additional sensors.

The first strategy motivates the need for models capable of removing as much systematic variability from the data as possible. The next section describes a framework combining models and data to detect shifts.

2.7 A model-based detection framework

The discussion in the previous sections motivates the need for statistical model-based detection methods. Figure 2-8 illustrates a framework that can be used to detect shifts in atmospheric CO_2 data given models and data. Several questions arise:

- Which model(s) should be used given a set of plausible models?
- What methods should be used to estimate the model parameters?
- What type of statistical test should be used?

The ingredients required by the framework are (1) models, (2) estimation and model selection methods, and (3) detection methods. These ingredients will are discussed in the following sections.

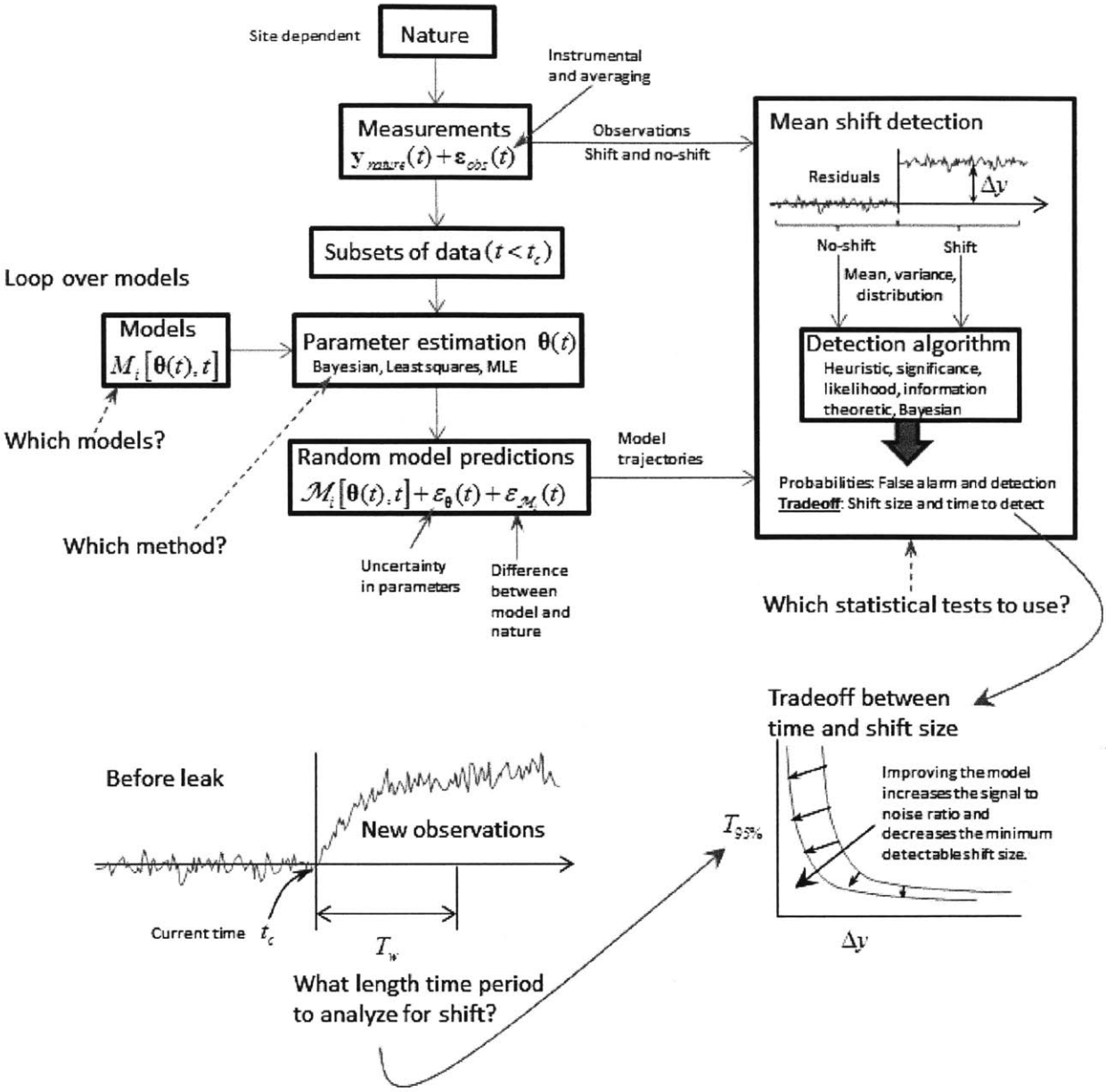

Figure 2-8: Model-based detection framework.

2.7.1 Models and forecasts

Any **model** $\mathcal{M}_i$ is an approximation to nature and will be prone to error. Thus,

$$y_{nature}(t) = y(\boldsymbol{\theta}, t|\mathcal{M}_i) + \varepsilon_{\mathcal{M}_i}(t) \tag{2.11}$$

where $\boldsymbol{\theta}$ is the vector of model parameters and $\varepsilon_{\mathcal{M}_i}$ is an unobservable error due to the model's inability to exactly capture the natural processes affecting the data. We call this error the **misfit error**. Misfit errors are model-dependent. The better the model, the smaller misfit errors will tend to be.

In practice, misfit error is unobservable: only the difference between the model and the data is observable. The data are related to the model predictions by

$$y(t) = y(\boldsymbol{\theta}, t|\mathcal{M}_i) + \varepsilon_{\mathcal{M}_i}(t) + \varepsilon_{obs}(t) \tag{2.12}$$

Given a particular value for $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$, the difference

$$e(t) = y(t) - y(\hat{\boldsymbol{\theta}}, t|\mathcal{M}_i) \tag{2.13}$$

is a series of **residuals**. Residuals are not random. They are specific outcomes of **model errors** defined by

$$\varepsilon(t) = \varepsilon_{\mathcal{M}_i}(t) + \varepsilon_{obs}(t) \tag{2.14}$$

Because of uncertainty in the parameters ($\boldsymbol{\theta}$) and uncertainty in the observing process (ε_{obs}), model errors can be treated as random variables. The scale of the uncertainty in model errors depends on both misfit error (that is, the chosen model) and on the measuring process. This scale can be estimated using the residuals:

$$\begin{aligned} \text{Var}\{\varepsilon(t)\} &\approx \text{Var}\{e(t)\} \\ &= \text{Var}\{\varepsilon_{\mathcal{M}_i}(t)\} + \text{Var}\{\varepsilon_{obs}(t)\} \\ &= \sigma^2_{\mathcal{M}_i} + \sigma^2_{obs} \\ &= \sigma^2_e \end{aligned} \tag{2.15}$$

As the model improves, $\sigma^2_{\mathcal{M}_i}$ will tend to decrease. This decrease is reflected in the magnitudes of residuals.

Future observations are subject to uncertainty in the model parameters. Since the data are uncertain, the model parameters in $\boldsymbol{\theta}$ will also be uncertain. When a single value of $\boldsymbol{\theta}$ is used to forecast CO_2, the future observations can thus be written

$$y(t) = y(\hat{\boldsymbol{\theta}}, t|\mathcal{M}_i) + \varepsilon_{\boldsymbol{\theta}}(t) + \varepsilon_{\mathcal{M}_i}(t) + \varepsilon_{obs}(t) \tag{2.16}$$

where $\varepsilon_{\boldsymbol{\theta}}$ indicates the error associated with the true $\boldsymbol{\theta}$ being different from $\hat{\boldsymbol{\theta}}$. We call $\varepsilon_{\boldsymbol{\theta}}(t)$ the **forward error** since it arises when uncertainty is propagated forward through the model. Forward errors will be model dependent, although this dependence has not been explcitly written in the above equation.

The variance of the forecast is given by

$$\begin{aligned} \text{Var}\{y(t)\} &= \text{Var}\{\varepsilon_\theta(t)\} + \text{Var}\{\varepsilon_{\mathcal{M}_i}(t)\} + \text{Var}\{\varepsilon_{obs}(t)\} \\ &= \sigma_\theta^2 + \sigma_{\mathcal{M}_i}^2 + \sigma_{obs}^2 \end{aligned} \tag{2.17}$$

Figure 2-9 illustrates the propagation of uncertainty in forming forecasts. Future observations can be modeled as

$$y(t) = y(\hat{\boldsymbol{\theta}}, t|\mathcal{M}_i) + \varepsilon_\theta(t) + \varepsilon_{\mathcal{M}_i}(t) + \varepsilon_{obs}(t) \tag{2.18}$$

If the best fit forecast $y(\hat{\boldsymbol{\theta}}, t|\mathcal{M}_i)$ is subtracted from both sides, then future residuals should behave according to

$$\varepsilon(t) = \varepsilon_\theta(t) + \varepsilon_{\mathcal{M}_i}(t) + \varepsilon_{obs}(t) \tag{2.19}$$

Therefore the shift detection problem can equally be viewed as the problem of detecting a mean shift in an uncertain forecasted series of residuals. The uncertainties due to modeling and the observation procedure remain the same, at least in magnitude.

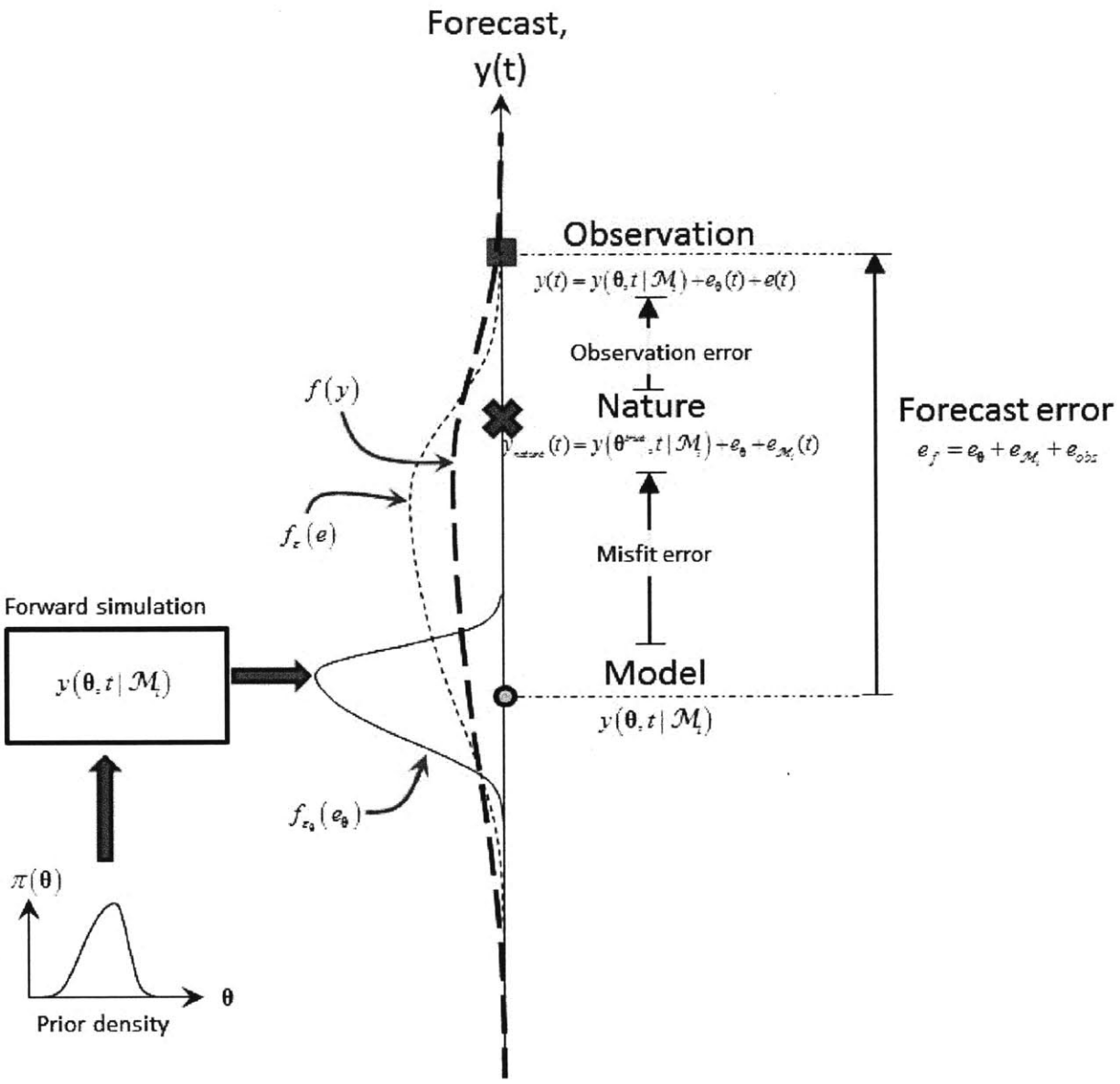

Figure 2-9: Forecasting CO_2.

Quantifying uncertainties in future CO_2 data Practically, this suggests the following procedure. Start by proposing a model $\mathcal{M}$. Next, estimate the model parameters. For example, least squares or Bayesian inference can be used to obtain means ($\hat{\boldsymbol{\theta}}$) and variances ($\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}$). The scale of forward uncertainty ($\hat{\sigma}^2_{\boldsymbol{\theta}}$) can be assessed by measuring the spread of random trajectories simulated using $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}$. The scale of observational uncertainty ($\hat{\sigma}^2_{obs}$) can be measured using (2.5). The scale of misfit uncertainty can be obtained as the difference between the variance of the residuals ($\hat{\sigma}^2$) and the variance of the observational errors ($\hat{\sigma}^2_{obs}$), as in $\hat{\sigma}^2_{\mathcal{M}} = \hat{\sigma}^2 - \hat{\sigma}^2_{obs}$

The uncertainty in the forecast is captured by the sum of the variances

$$\begin{aligned} \hat{\sigma}^2_{forecast} &= \hat{\sigma}^2_{\boldsymbol{\theta}} + \hat{\sigma}^2_{\mathcal{M}} + \hat{\sigma}^2_{obs} \\ &= \hat{\sigma}^2_{\boldsymbol{\theta}} + \hat{\sigma}^2 \end{aligned} \tag{2.20}$$

$\hat{\sigma}^2_{forecast}$ measures the uncertainty associated with individual future observations. If the future observations are independent Gaussian random variables with mean y and variance $\hat{\sigma}^2_{forecast}$, then the sample mean of N future observations will have a 95% confidence interval given by

$$y(\hat{\boldsymbol{\theta}}, t|\mathcal{M}_i) \pm 1.96\frac{\hat{\sigma}_{forecast}}{\sqrt{N}} \tag{2.21}$$

The mean shift detection problem is essentially the problem of comparing confidence intervals like this one. If they overlap, then the hypothesis that a shift has occurred cannot be rejected. If they do not, then the no-shift hypothesis is rejected in favor of the hypothesis that a shift has occurred.

2.7.2 Formalizing the shift detection problem

The shift detection problem is the problem of choosing between multiple hypotheses. Given N observations $[y(t_1), ..., y(t_N)]$ and a model $\mathcal{M}$ for making predictions, the question is, is a shift in the mean present? The data are related to the model by

$$y(t) = \mathcal{M}(\boldsymbol{\theta}, t) + \Delta y \cdot \delta(t - t_s) + \varepsilon_{\mathcal{M}}(t) + \varepsilon_{obs}(t) \tag{2.22}$$

where $\boldsymbol{\theta}$ is the vector of detrending model parameters, Δy and t_s are shift parameters which in the following will be grouped into the vector $\boldsymbol{\eta}$, and where $\delta(t - t_s)$ is given by

$$\delta(t - t_s) = \begin{cases} 0 & t < t_s \\ 1 & t \geq t_s \end{cases} \tag{2.23}$$

$\varepsilon_{\mathcal{M}}$ is the error associated with the detrending model's approximation of nature. ε_{obs} is the sum of the errors from instrumentation and averaging.

If no shift exists then we adopt the convention of setting $t_s = t_0$ and the mean (Δy) should be close to zero. Otherwise a shift occurs and Δy is non-zero.

The different possible events represent competing hypotheses, each with its own associated model $\mathcal{L}_i$ ($i = 0, 1, ..., N$) of the data. $\mathcal{L}_0$ represents the no-shift hypothesis

for which $t_s = t_0$ and $\Delta y = 0$. $\mathcal{L}_{i>0}$ represents the $t_s = t_i$ hypothesis, for which $\Delta y = 0$ until $t = t_s$ at which time it shifts to a new value $\Delta y \neq 0$. To be more concise, the no-shift hypothesis has unknown Δy_0, which when estimated should be close to zero. The single-shift hypothesis has unknowns t_s, Δy_0 (which should be near zero), and Δy_1 (the shift size, which should be positive). Here, the unknowns for a given hypothesis are lumped into the vector $\boldsymbol{\eta}$. The value predicted by the complete model depends on both the detrending model parameters as well as the shift parameters, and can be denoted $y(\boldsymbol{\theta}, \boldsymbol{\eta}, t)$. In the following, it is assumed the detrending parameters are held fixed while the detection algorithm proceeds. The predictions are still dependent on the particular model chosen for detrending, but for conciseness they will be referred to by $y(\boldsymbol{\eta}, t)$.

The formulation just described can be extended without loss of generality to the case where multiple shifts are possible in the observations. However, as argued shortly, consideration of multiple shifts is not important for leak detection. Therefore, algorithms are presented in the context of a single possible shift. This maintains focus on the leak detection problem and has the added benefit of keeping notation simple.

2.8 Statistical approaches to detection

2.8.1 Approaches

Several different approaches are available for determining if, when (t_s) and how big (Δy) a shift in the observations has occurred. We distinguish between the following approaches:

- **Heuristics.** These are based on measures of distance involving the predictions and the observations. Different formulations are possible. One class of formulations is based on the distance between the observations and the predictions. An example is the sum of squared residuals:

$$RSS = \sum_{j=1}^{n_{obs}} [y(t_j) - \mathcal{M}(\boldsymbol{\eta}, t_j)]^2 \tag{2.24}$$

  The shift detection procedure would be as follows. First, hypothesize a time of shift t_s. Estimate Δy for the chosen shift time as the mean of the samples for which $t \geq t_s$. Compute the sum of squared residuals, let us denote it $RSS(t_s)$. Do this for all possible shift times. Choose the shift time for which RSS is minimized. If the shift time obtained is $t_s = t_0$, then there is no shift.

  Other criteria can be substituted for RSS. Examples include the maximum absolute deviation ($\max_j y(t_j) - \mathcal{M}(\boldsymbol{\eta}, t_j)$) and the sum of absolute deviations ($\sum_j y(t_j) - \mathcal{M}(\boldsymbol{\eta}, t_j)$).

  Another class of criteria are based on zonation techniques [25, 26]. These techniques split the series into two zones around the hypothesized shift time. The first zone has data $\boldsymbol{y}_1 = [y_1 : y_{s-1}]$ associated with it. The second zone comprises

$\boldsymbol{y}_2 = [y_s : y_N]$. Each zone has its own mean value, $\bar{\boldsymbol{y}}_1$ and $\bar{\boldsymbol{y}}_2$ respectively. The mean of all of the samples has some other value $\bar{\boldsymbol{y}}$. Based on these values and the observations, it is possible to define measures of "within-zone" variability as

$$W = \frac{1}{N-2} \sum_{z=1}^{2} \sum_{k=1}^{n_z} (\boldsymbol{y}_z(k) - \bar{\boldsymbol{y}}_z)^2, \tag{2.25}$$

where the factor of 2 in the denominator is the number of zones; and "between-zone" variability as

$$B = \sum_{z=1}^{2} n_z \left(\bar{\boldsymbol{y}}_z - \bar{\boldsymbol{y}}\right)^2, \tag{2.26}$$

The criterion used to determine the best shift time is given by

$$R = \frac{B - W}{B} \tag{2.27}$$

The value of R will range between 0 and 1. The best shift time is the one that obtains R closest to one. The motivation is to choose the shift time that split the series into widely separated zones, given the natural variability in the data. There is a caveat in applying the above zonation procedure: it assumes there are two distinct zones in the data. The zonation procedure is only applicable when you already know a shift is present in the data. Such a procedure is obviously not useful for leak detection.

A second, more concerning issue associated with the procedures described above is their inability to provide theoretically sound stopping points. In general, the more "zones" the better the value of the criterion obtained. The decision to add an additional shift time into the model is made based on a heuristic rule defined by the researcher. Clearly this is unsatisfactory for leak detection because it means different researchers will come to different conclusions, with the detection criteria providing no objective way to reconcile their different points of view.

- **Significance.** Tests of statistical significance are the most widely used methods for hypothesis testing. The procedure is to propose a null hypothesis (in this case, $\mathcal{L}_0$) and then reject or fail to reject it baesd on the value of a test statistic. The test statistic is usually translated into a p-value. p-values are often misinterpreted. They represent the probability of observing a more extreme test statistic value under the null hypothesis. Before the test, the researcher sets a threshold called the level of significance (α). If the p-value is smaller than α, then the null hypothesis is rejected. Otherwise no decision can be made between the competing hypotheses.

  An extremely common detection method based on the significance paradigm is to compare 95% sample mean confidence intervals. The $100 \times (1-\alpha)\%$ sample mean confidence interval of a small number of samples (N) is given by

$$\overline{\Delta y_{1:N}} \pm t_{(\alpha/2, N-1)} \cdot \hat{\sigma}/\sqrt{N} \tag{2.28}$$

where $\hat{\sigma}^2$ is the sample variance of the residuals $\Delta y(t_i)$ $(i = 1, ..., N)$ and $t_{()}$ is the upper critical value of the t distribution with $N - 1$ degrees of freedom. As N increases, the asymptotic value of $t_{()}$ is 1.96 when α=0.05 because the t distribution approaches a standard normal distribution as $N \to \infty$. Thus, 95% sample mean confidence intervals are often computed as $1.96\hat{\sigma}/\sqrt{N}$ and for rough calculations $2\hat{\sigma}/\sqrt{N}$ is sometimes used.

Other examples of tests include the Z-test and different versions of Student's t-test. Table 2.1 summarizes these methods. Standard tests like t-tests assume the residuals are independent and have identical, normal distributions, but nonparametric tests also exist that can be used to extend significance testing to non-normal distributions. One of the biggest setbacks in using significance tests is that they are unable to choose between multiple hypotheses: p-values cannot be used to rank hypotheses, they can only be used to reject them. Even in the simplest case, where $\mathcal{L}_{i>0}$ represents the possibility of a single shift being present, there are $N+1$ hypotheses to choose from. (N are from the N possible shift times and one is from the no-shift hypothesis.) Therefore, unless the shift time is known, which in our case it is not, standard significance tests are poorly suited to leak detection.

- **Likelihood.** These methods rely upon the value of the likelihood function. Let us assume that the problem is to detect a shift in the mean of normally distributed observations all having the same variance σ^2. The portion of the log likelihood function of interest would be given by

$$L(t_s, \Delta y) = -\sum_{j=1}^{N} [y(t_j) - \mathcal{M}(\boldsymbol{\eta}, t_j)]^2 \tag{2.29}$$

which is just the opposite of the sum of squared residuals. Maximizing the above likelihood function results in "optimal" estimates for t_s and Δy. Hawkins [27] shows this problem can be solved with a dynamic programming approach. As with the heuristic algorithms presented earlier, however, a sound method of determining the number of shifts present, if any, does not exist: the optimal value of the likelihood function can be increased by introducing additonal shift times into the model.

Lavielle [28] introduces a penalized version of the likelihood function. His approach for detecting a single shift in observations that differ from the model by Gaussian white noise (see equation (15) in the paper) would entail minimizing

$$U(t_s) = RSS(t_s) + \beta \cdot [1 - \delta(s, 0)] \tag{2.30}$$

where β (not to be confused with $\boldsymbol{\beta}$) controls the eagerness of the algorithm to declare shifts and $\delta(i, 0)$ is the Dirac delta function. The dilemma is what value to assign to β. Large values will lead to decisions in favor of no-shift, while small values will have the opposite effect. Lavielle's approach is qualitatively appealing because it captures the tradeoff between goodness-of-fit and model

Table 2.1: Significance tests for mean shift detection

Test	Formula[b]	Assumptions
Z-test	$\mathcal{T} = \frac{\overline{\Delta y_{1:N}}}{\sigma/\sqrt{N}}$	– Model errors Gaussian and independent. – Population variance known, does not change. – Shift has occurred before t_1.
Student's t-test	$\mathcal{T} = \frac{\overline{\Delta y_{1:N}}}{\hat{\sigma}/\sqrt{N}}$	– Model errors Gaussian and independent. – Population variance unknown, does not change. – Shift has occurred before t_1.
Two-sample t-test	$\mathcal{T} = \frac{\overline{\Delta y_{1:s-1}} - \overline{\Delta y_{s:N}}}{\hat{\sigma}/\sqrt{N}}$	– Model errors Gaussian and independent – Population variance known, may change. – Tests for shift at time t_s, where $1 \leq s \leq N$.

[a] The leak detection problem is to detect a shift in the residuals. The mean should be zero before a leak arises.

[b] Notation used:
$\Delta y(t_i) = y(t_i) - \mathcal{M}(\boldsymbol{\eta}, t_i)$ is the residual, or model error, at time t_i.
$\underline{\Delta y_{m:n}} = [\Delta y(t_m), ..., \Delta y(t_n)]$ is a vector of residuals.
$\overline{\Delta y_{m:n}} = \frac{1}{m-n+1} \sum_m^n \Delta y_i$.
σ is the population variance of the residuals.
$\hat{\sigma}$ is a measure of sample variance of the residuals. Its exact definition depends on the test.

complexity. However, the absence of a strong statistical rationale for choosing β means that this approach will still leave researchers unable to reconcile differing views when they arise.

- **Information theoretic.** A number of information criteria exist for choosing between hypotheses like the ones we are considering [29, 30, 31, 32]. The main ones are Akaike's criterion (AIC and AICc), Schwarz's Bayesian criterion (BIC), and Spiegelhalter's Deviance information criterion (DIC).

  The starting point for information based criteria is the Kullback-Leibler information loss

$$I(G,g) = \int G(\boldsymbol{y}_{obs}) \ln \frac{G(\boldsymbol{y}_{obs})}{g(\boldsymbol{y}_{obs}|\boldsymbol{\eta})} d\boldsymbol{y}_{obs} \tag{2.31}$$

  Here, G represents the true distribution of the observations implied by nature and g the likelihood implied by our model. For example, if we assume the data at time t are Gaussian with mean $y(\boldsymbol{\eta}, t)$ then $g \propto \prod_j \exp\{-[y_{obs,j} - \mathcal{M}(\boldsymbol{\eta}, t_j)]^2/2\sigma^2\}$. $\boldsymbol{\eta}$ is restricted to continuous parameters, so it can include physical model parameters in addition to the shift parameters (Δy's) of interest. Physical model parameters would be included in $\boldsymbol{\eta}$ if the model parameters are allowed to adjust to the new data. Otherwise the physical parameters are held fixed at their prior values and $\boldsymbol{\eta}$ may just include shift parameters.

  The Kullback-Leibler information loss represents the information lost when g is used to approximate G. The problem is that G is never known in practice. The way around this is to rewrite K-L information loss as

$$I(G,g) = \int G(\boldsymbol{y}_{obs}) \ln G(\boldsymbol{y}_{obs}) d\boldsymbol{y}_{obs} - \int G(\boldsymbol{y}_{obs}) \ln g(\boldsymbol{y}_{obs}|\boldsymbol{\eta}) d\boldsymbol{y}_{obs} \tag{2.32}$$

  The first term does not depend on the parameters and thus can be dropped from further consideration. Information theoretic criteria are based on estimates of the second integral in (2.32).

  Akaike's information criterion (AIC, or a modified version AICc) is very commonly encountered. AIC is based on an asymptotic estimate of $\int G(\boldsymbol{y}_{obs}) \ln g(\boldsymbol{y}_{obs}|\boldsymbol{\eta}) d\boldsymbol{y}_{obs}$. It is defined as

$$AIC = 2k - 2 \ln \ell(\boldsymbol{\eta}|\boldsymbol{y}_{obs}) \tag{2.33}$$

  where k is the number of parameters in $\boldsymbol{\eta}$ ($k = 1$ parameter, Δy_0, if there is no shift; and $k = 2$ parameters, Δy_0 and Δy_1, if there is one), and ℓ is the likelihood function associated with the observations. For example, if the data are normally distributed with Gaussian variance about the model predictions then

$$AIC = 2k + 2 \ln \left[\frac{RSS(\boldsymbol{\eta})}{2\sigma^2} \right] \tag{2.34}$$

  Since $\boldsymbol{\eta}$ cannot include discrete parameters, the shift time t_s must be incremented manually and AIC is calculated for each. Hypotheses ($\mathcal{L}_i$) resulting in smaller AIC values are more favorable. The Gaussian case above highlights the

trade-off between model complexity (which grows with k) and goodness-of-fit (RSS). Since AIC only considers part of the K-L information, the magnitude of AIC obtained for a particular model is without meaning. However, interpretation is possible through consideration of differences in AIC's obtained for different models. Burnham and Anderson [33, 34] demonstrate how AIC values can be translated into "Akaike weights", which from a Bayesian standpoint can be interpreted as posterior odds. This is a major strength of information theoretic methods: they provide a statistically sound basis for determining when and how big a shift occurred.

- **Bayesian.** Bayesian methods to shift detection would base their decisions on posterior probabilities, or ratios of posterior probabilities, of the hypotheses given the data. Examples of Bayesian methods of model selection include Bayes factors, Markov switching models and supermodeling (see [35, 36] for descriptions of each). Bayesian methods require prior distributions for the unknowns, and are capable of handling all questions regarding the number, size and times of shifts. They tend to be more computationally intensive than information theoretic methods but provide a similarly sound statistical basis for deciding whether, when and how big a shift has occurred.

2.8.2 Dealing with multiple shifts

Many of the criteria introduced above are applicable to the case where multiple shifts are to be detected. In some cases, the criteria we reference were even created specifically for the purpose of detecting multiple shifts. In detecting leaks from geological sequestration, we are really only interested in two shifts: (1) is there an unexpected increase in CO_2, and (2) after mitigation, does CO_2 decrease. From a purely mathematical standpoint, these problems are equivalent. Detecting an increase of Δy should be just as difficult as detecting the opposite. From a practical standpoint, we are really just interested with the first case.

For a different problem, which is estimating leakage rate and location by inverting perturbations in CO_2 concentration to get fluxes, the key is obtaining accurate estimates of the shift as a function of time ($\Delta y(t)$). We could think about using multiple shift detection methods to approximate the true CO_2 enhancement from leakage. This approach would be useful when a relatively large time period is being considered, in which multiple shifts are likely to have occurred. A simpler approach is to slide a smaller window across the data of interest and look for individual shifts one at a time. Each time a shift is detected, the mean of the residuals after the determined shift time can be used to estimate the CO_2 perturbation. By sliding the window across the data, a smooth curve can be produced for $\Delta y(t)$. The ability to resolve this smooth curve $\Delta y(t)$ will depend on the size of the window required to detect individual shifts at the desired level of statistical significance. If the size of this window can be reduced, then more sudden the changes in $\Delta y(t)$ should be resolvable. Choosing the appropriate window size, as discussed in Chapter 3, is an important step in being able to estimate shifts in atmospheric CO_2 due to leakage.

2.8.3 Normality and autocorrelation

Most methods rely on the assumptions that forecasting errors are normally distributed and independent. If errors are non-Gaussian, then Bayesian methods are needed for detection. Use of non-Bayesian detection methods when uncertainties are non-Gaussian will lead to loss of control over false alarm rate. When errors are autocorrelated, it is important to use estimation methods capable of handling the correlation. Common methods of estimation like ordinary least squares do not account for autocorrelation and tend to result in overconfidence parameter estimates whose consequence is increased false positive rates. Normality and autocorrelation are discussed in more detail next.

2.9 Normality

Figure 2-10 shows the distribution of hourly CO_2 data from Harvard Forest, a site which will be studied in more detail later. The site is subject to local and regional sources and sinks of biological origin. The data exhibit substantial variability due to these sources and sinks. From the figure it is clear that the data are not normally distributed. This can be confirmed using normality tests.

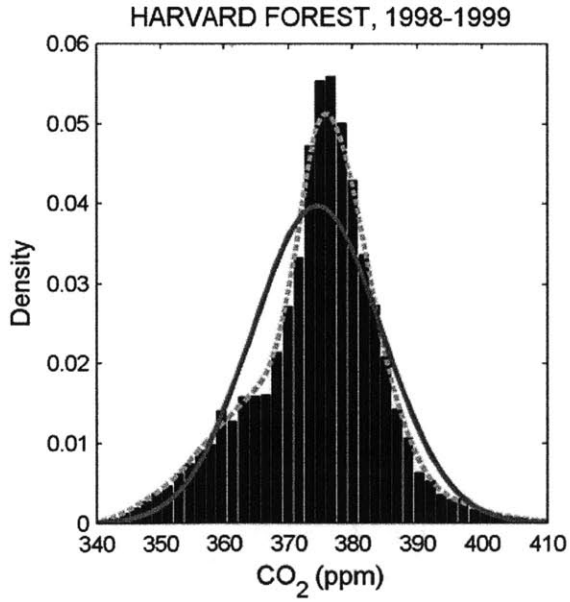

Figure 2-10: Distribution of hourly CO_2 data from Harvard Forest between Jan 1, 1998 and Dec 31, 1999. The blue bars are a normalized histogram of the data. Normal (solid red) and empirical (dashed green) densities based on the data are also plotted.

2.9.1 Testing normality

Most normality tests are qualitatively similar insofar as they both involve comparisons between expected and observed probability densities. When choosing between multiple distributions for the same data, information theoretic results like K-L distance (for an example see Burnham and Anderson [33]) are best used because p-values are not intercomparable. Significance tests are useful when the goal is simply to determine whether the residuals are Gaussian.

There are many ways to test normality. Exploratory approaches include normal probability plots[3] and plotting normal density fits alongside normalized histograms as done in Figure 2-10. These types of methods rely on the eye of the user to distinguish whether normality is reasonable. Many quantitative alternatives exist. The various approaches follow along the same lines as the detection approaches outlined earlier. The most statistically rigorous approach of which the author is aware is due to Burnham and Anderson [33], who demonstrate the use of information criteria to ranking several proposed distributions when the "true" distribution is unknown. More standard approaches include those based on classical and nonparametric significance tests. Numerous tests of this type are available: for a survey of several methods, the reader is referred to D'Agostino and Stephens [37]. Examples of statistical significance tests for normality include the chi-square, Jarque-Bera, Kolmogorov-Smirnov, and Lilliefors tests. We used these tests in our analysis, a choice motivated by the fact that routines were already available for them in MATLAB. Brief descriptions of each test are provided below.

Chi-square goodness-of-fit test (chi2gof)

Chi-square normality test would compare the histogram of the e_t's to that which would be expected if the observations were normally distributed with mean and variance given by the sample mean and variance. The test statistic is

$$\mathcal{T} = \sum_{i=1}^{n_{bins}} (O_i - E_i)^2 / E_i \tag{2.35}$$

where O_i and E_i are the observed and expected numbers of candidate values falling into bin i. The null hypothesis is that the observations come from a normal distribution parameterized by the sample mean and variance. The test statistic has an approximate chi-square distribution with $n_{bins} - 3$ degrees of freedom under the null hypothesis.

[3]Probability plots work by plotting the percentiles of the observations against the percentiles from a normal distribution with mean and variance equal to the sample mean and variance of the observations. If the curve deviates from a straight line, for example because of fat tails or skew, then the normality hypothesis is rejected.

Jarque-Bera test (jbtest)

The test proposed Jarque and Bera [38] by is a two-sided goodness-of-fit test based on the sample skewness and sample kurtosis. The test statistic is

$$\mathcal{T} = N/6 \left[\left(\sqrt{b_1} \right)^2 + \frac{(b_2 - 3)^2}{4} \right] \tag{2.36}$$

where

$$\sqrt{b_1} = \hat{\mu}_3 / \hat{\mu}_2^{3/2} \tag{2.37}$$

is the sample skewness and

$$b_2 = \hat{\mu}_4 / \hat{\mu}_2^2 \tag{2.38}$$

is the sample kurtosis and $\hat{\mu}_j = \Sigma (e_t - \bar{e})^j / N$ and $\bar{e} = \Sigma e_t / N$. When the sample size (N) is sufficiently large, the test statistic has an approximate chi-square distribution with two degrees of freedom under the null hypothesis. In MATLAB, jbtest enables the Jarque-Bera test to be applied for smaller sample sizes ($N < 2{,}000$) using tables based on Monte Carlo simulations. The null hypothesis is that the samples come from a normal distribution with unknown mean and variance. The alternative is that the they come from a distribution in the Pearson family other than the normal distribution.

Kolmogorov-Smirnov test (kstest)

This is a distribution free test. The one-sample version of the test can be used to test against the null hypothesis that the samples have a standard normal distribution with unknown mean and variance. The test statistic is based on the difference between the emipirical cumulative distribution function of the samples and that for a standard normal distribution.

Lilliefors test (lillietest)

The null hypothesis is that the samples come from a normal distribution with unknown mean and variance. The alternative is that the samples belong to a non-normal distribution. The test statistic is defined just like the test statistic for the K-S test, except that the empirical cdf is compared to the cdf of a normal distribution with mean and variance determined by the sample mean and variance of the observations. Lilliefors test is a two-sided goodness-of-fit test.

2.9.2 Results for the Harvard data

Applying the Jarque-Bera test to the Harvard data in Figure 2-10 results in rejection of normality at the 5% level of significance. The Jarque-Bera test statistic is 580, which far exceeds the 5% threshold value of 5.99. The test statistic corresponds to a P-value less than 0.001. The chi-square and Kolmogorov-Smirnov tests also reject normality at the 5% level of significance.

2.9.3 Implications of non-Gaussian distributions

However, the normality assumption is often not tested. This is especially a problem when there are few samples to work with, in which case the Central Limit theorem may not apply. For example, consider the Z-test. The Z-test compares the mean of the samples, $\bar{y}_{obs}$, to a known value. If the observations are normally distributed, then the sample mean will be also. What if the observations are not normally distributed? The Central Limit theorem suggests that for a *large enough number* of observations, the sample mean will be Gaussian. However, if insufficient observations are considered then the sample mean need not be Gaussian. This has obvious implications for our ability to detect. In the case of the Z-test, the test statistic will not be standard normal which means that the actual rate of false alarm may be higher or lower than that prescribed by the researcher. The implication is that violation of normality implies a loss of control over the performance of the detection algorithm. Bayesian methods should be used when the model residuals are not normally distributed because they can handle any type of distribution.

2.10 Parameter estimation

Substantial fluctuations are naturally present in the data. Figure 2-11 shows the empirical distribution of the fluctuations alongside a Gaussian fit after a smooth seasonal trend has been removed from the data. The difference between the two suggests the need for Bayesian methods.

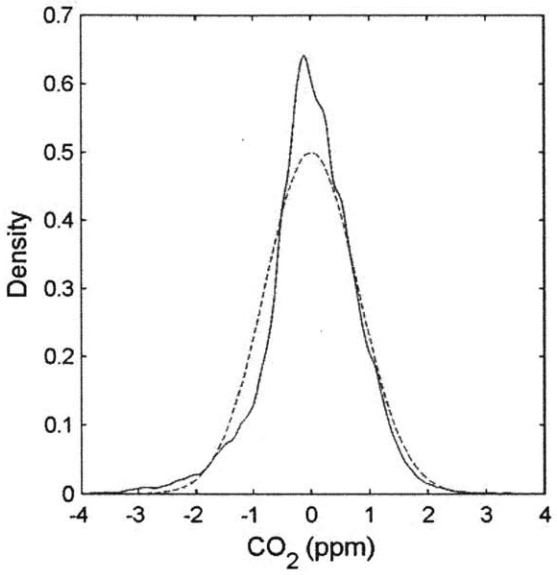

Figure 2-11: Distribution of fluctuations in CO_2. Normal (red, dashed), and empirical (blue).

2.10.1 The problem

Bayesian inference refers to the use of Bayesian methods for making inferences about unknowns from data. From the Bayesian perspective - in contrast with that of the frequentist - the unknowns are uncertain. Bayes' theorem provides the basic tool needed to adjust certainty in these unknowns after receiving new data y.

Bayesian inference is based on Bayes' theorem. Bayes' theorem is a rule for updating beliefs given data when degree of belief, or certainty, is measured by probabilities. The discrete version is applicable to model selection

$$P(\mathcal{M}|y) = \frac{f(y|\mathcal{M})P(\mathcal{M})}{f(y)} \tag{2.39}$$

$P(\mathcal{M})$ is the prior probability assigned to model $\mathcal{M}$. After observing new data y, the new probability is the posterior $P(\mathcal{M}|y)$. The theorem is useful when there are multiple competing models. The bigger the probability associated with a model, the more likely it is to be the target model. $f(y)$ is the marginal density of the data, the distribution of the observations averaged over all possible models. $f(y|\mathcal{M})$ is the density function of the data conditioned on model $\mathcal{M}$.

The continuous version of Bayes' theorem is useful for estimating the values of unknown, continuous model parameters. It states that

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \tag{2.40}$$

Prior knowledge of θ, which exists before the observations are available, is contained by the prior density $\pi(\theta)$. Posterior knowledge is encapsulated by the posterior density $\pi(\theta|y)$. The third term appearing in (2.40) is the likelihood $f(y|\theta)$, which as the name describes measures the likelihood of observing y given a value for θ. The integral in the denominator of (2.40) is a normalization constant that ensures the posterior density is a proper density function.

In practice, we are interested not only in the posterior density itself but in some of its features. Inference results are often summarized in terms of features like the mean, mode, and $(1 - \alpha)$ confidence intervals. Most features can be expressed as weighted arithmetic averages of the form

$$E\left[f(\theta)|y\right] = \frac{\int f(\theta)\pi(\theta|y)d\theta}{\int \pi(\theta|y)d\theta} \tag{2.41}$$

The primary goals in Bayesian inference are thus solving (2.40) and evaluating expectations like (2.41). The biggest obstacle is solving (2.40) because the integral in the denominator is usually difficult to evaluate. A special case when analytical evaluation is possible occurs when the prior density has conjugacy with the family of the likelihood. In this case the posterior density will have the same functional form as the prior density and the integral can be ignored. An alternative which is only sometimes appropriate is to use an approximation method. Some approximation methods, like Laplace appriximation, can modify the problem in a way that makes it analytically tractible. When analytical evaluation is not possible, one must resort to numerical integration or Monte Carlo integration. Numerical integration methods like quadrature fail when θ has a large number of components. Monte Carlo integration methods are flexible enough to handle problems where numerical integration fails. The tradeoff is that they can be challenging to implement.

When, for a given likelihood $f(y|\theta)$, the prior density can be chosen so that the posterior density belongs to the same family, then the prior density family is said to have **conjugacy** with the family of the likelihood. If conjugacy is present, then the posterior density's parameters can be determined from the observations y and the prior density's parameters. Conjugacy is important because it makes computationally efficient recursive estimators like recursive least squares and the Kalman filter possible. These are useful for leak detection from geological sequestration, where N new observations will be considered at a time and if no leak is detected will need to be assimilated into the baseline model.

2.10.2 Ordinary least squares

Ordinary least squares is one special case where the posterior density can be obtained analytically. The assumptions are:

- Observables are linearly dependent on the unknowns and are subject to additive Gaussian noise.
- The unknowns (θ) do not vary in time.
- In ordinary least squares, as opposed to generalized least squares, the noise is independent and identically distributed (i.i.d.) Here, we assume the noise is $N(0, \sigma^2)$, or zero mean with variance σ^2, where σ^2 is known. In general, the noise variance can be estimated alongside the other unknowns. In least squares this is usually done in a two step fashion; in theory Bayesian inference can estimate (θ, σ^2) in a single pass. The Bayesian approach would require us to incorporate prior knowledge about σ^2.
- There is zero prior knowledge about the values of the unknowns.

Practically, these assumptions specify a model wherein

$$\begin{aligned} y_t &= \sum_{i=1}^{n_\beta} x_{i,t}\beta_i + \varepsilon_t \\ f(y_t|\beta_1, \beta_2, ..., \beta_{n_\beta}) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[y(t) - \mathcal{M}(\boldsymbol{\beta}, t)]^2}{2\sigma^2}\right\} \\ \pi(\beta) &= \text{a constant (an improper density)} \end{aligned} \tag{2.42}$$

The first two equations in (2.42) can be simplified using matix notation. Let

$$\begin{aligned} \boldsymbol{y} &= [y_1, y_2, ..., y_{nobs}]^T \\ \boldsymbol{x}_t &= [x_{t,1}, x_{t,2}, ..., x_{t,n\beta}]^T \\ \boldsymbol{X} &= [\mathbf{x}_1, \mathbf{x}_2, ..., \boldsymbol{x}_{nobs}]^T \\ \boldsymbol{\beta} &= [\beta_1, \beta_2, ..., \beta_{n\beta}]^T \\ \boldsymbol{\varepsilon} &= [\varepsilon_1, \varepsilon_2, ..., \varepsilon_{nobs}]^T \end{aligned}$$

In matrix form, the assumption of i.i.d. Gaussian noise can be written

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{nobs})$$

The model in (2.42) becomes

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ f(\boldsymbol{y}|\boldsymbol{\beta}) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n_{obs}} \exp\left[-\frac{[\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}]^T[\boldsymbol{y}_{obs} - \boldsymbol{X}\boldsymbol{\beta}]}{2\sigma^2}\right] \\ \pi(\boldsymbol{\beta}) &= \text{a constant (no change from before)} \end{aligned} \tag{2.43}$$

It can be shown that the posterior density arising from the model in (2.43) is Gaussian with mean and variance

$$\begin{aligned}\bar{\boldsymbol{\beta}} &= E\left[\boldsymbol{\beta}|\boldsymbol{y}\right] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\end{aligned} \tag{2.44}$$

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{\beta}} &= E\left[(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^T|\boldsymbol{y}\right] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned} \tag{2.45}$$

Specifically, the posterior density is

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{1}{(2\pi)^{n_\beta/2}(\det \boldsymbol{V}_{\boldsymbol{\beta}})^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^T\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\right] \tag{2.46}$$

It turns out that $\bar{\boldsymbol{\beta}}$ and $\boldsymbol{V}_{\boldsymbol{\beta}}$ are the same mean and variance one obtains as the solution to the ordinary least squares problem,

$$\min_{\boldsymbol{\beta}}[\boldsymbol{y}_{obs}-\boldsymbol{X}\boldsymbol{\beta}]^T[\boldsymbol{y}_{obs}-\boldsymbol{X}\boldsymbol{\beta}] \tag{2.47}$$

Thus, in the special case that the assumptions listed in the beginning are correct, ordinary least squares is a Bayesian inference method. Violation of any of the original assumptions means the posterior density will no longer be Gaussian and the ordinary least squares estimators in (2.44) and (2.45) no longer accurately capture the posterior knowledge about the unknowns.

2.10.3 Nonlinear least squares

Nonlinear least squares applies when the model relating the parameters and observed variables is nonlinear. Thus,

$$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{\theta}) + \boldsymbol{\varepsilon}. \tag{2.48}$$

The problem is that when the model is nonlinear, the posterior density of the parameters is almost guaranteed not to be Gaussian. This means that the confidence regions will tend not to be perfectly ellipsoidal. Nonlinearity also can affect the least squares value we obtain. There are two reasons for this. First, the when the Jacobian of the model relation $\boldsymbol{g}$, which we will call $\boldsymbol{J}$, becomes extremely ill conditioned, it can make finding the optimal value of $\boldsymbol{\theta}$ difficult. In ordinary least squares, the optimal value is reached in a single step by computing 2.44 and so ill conditioned $\boldsymbol{X}$ does not represent a problem expect possibly a numerical one for the inversion of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. In nonlinear least squares, however, findig the optimal value requires making an initial guess and then using an optimization algorithm to iteratively find the optimal solution. In this case, having ill conditioned $\boldsymbol{J}$ makes the contours of the least squares objective function narrow in certain dimensions, which can make finding the optimal solution extremely challenging. The second way in which nonlinearity affects our ability to find an optimal solution is due to the fact that multiple minima can exist. The consequence is that finding the *global* minimum can be a

challenge, especially when the objective function takes time to compute. Methods like simulated annealing exist that can be applied to finding global solutions to the nonlinear least squares problem. Let us overlook this technicality and assume that a solution has been found, since our primary interest here will be to demonstrate the effect nonlinearity has on the confidence regions for the least squares values. Two versions of confidence intervals can be found in the literature and both are described on pp. 97-98 of Seber and Wild [39]. (For a discussion that parallels that of Seber and Wild, the reader may find pp. 308-309 of Sen and Srivastava [40] useful. Also, pp. 807-818 of Press et al. [41] gives a sampling approach to obtaining confidence regions.) The two major alternative definitions of confidence regions for nonlinear least squares problems are described below:

1. *Asymptotic confidence regions.* This approach is the more common one. Given a level of significance α, a confidence region can be determined as the contour determined by
$$\left\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T \bar{\boldsymbol{J}}^T \bar{\boldsymbol{J}} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \leq p\hat{\sigma}^2 F^{\alpha}_{p,n-p}\right\} \tag{2.49}$$
where n is the number of observations, p is the number of systematic model parameters in $\boldsymbol{\theta}$, F^{α} is the upper critical value of the F distribution, $\hat{\sigma}^2$ is the sample variance of the residuals computed as $RSS(\bar{\boldsymbol{\theta}})/(n-p)$, and $\bar{\boldsymbol{J}}$ is the value of the Jacobian matrix obtained for $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$. The key idea is that (2.49) is based on a linear approximation to $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})$ which is only a good approximation in some small neigborhood of the least squares value $\bar{\boldsymbol{\theta}}$. Within this neighborhood, however, the model function is approximately linear, so the posterior distribution looks Gaussian and the confidence regions within this neighborhood can be well approximated by the ellipsoids given by (2.49). If (2.49) is good approximation to the actual posterior distribution of the parameters, then the posterior distribution can be represented using just the least squares estimate $\bar{\boldsymbol{\theta}}$ and the covariance matrix
$$\boldsymbol{V}_{\bar{\theta}} = \left(\bar{\boldsymbol{J}}^T \bar{\boldsymbol{J}}\right)^{-1} \hat{\sigma}^2 \tag{2.50}$$

2. *Exact confidence regions.* So-called "exact" confidence regions arise to handle the case where the linear approximation to $\boldsymbol{g}()$ is a poor one, given the level of significance of interest. (Note that by choosing a bigger level of significance, we focus on a smaller neighborhood around $\bar{\boldsymbol{\theta}}$ so for some big enough value of α we can expect the linear approximation to be a good one. However, we are usually interested in small α's, which in the case of nonlinear least squares can necessitate consideration of the confidence regions presented next.) Exact confidence regions are formed directly from the contours of the nonlinear least squares objective function:
$$\left\{\boldsymbol{\theta} : \Phi(\boldsymbol{\theta}) \leq \Phi(\bar{\boldsymbol{\theta}}) \left(1 + \frac{p}{n-p} F^{\alpha}_{p,n-p}\right)\right\} \tag{2.51}$$

where Φ is the least squares objective function defined by

$$\Phi(\boldsymbol{\theta}) = [\boldsymbol{y}_{obs} - \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})]^T [\boldsymbol{y}_{obs} - \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})] \tag{2.52}$$

Seber and Wild [39] demonstrate that this region is valid when n is large.

2.10.4 Forward sensitivity method

Sensitivity methods are useful for determining least squares estimates for parameters

$$\boldsymbol{\theta} = [\theta_j] \quad j = 1, 2, ..., p \tag{2.53}$$

when the observations are given by

$$\boldsymbol{y}(t) = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{\theta}, t) + \boldsymbol{\varepsilon}(t) \tag{2.54}$$

and the state variables $\boldsymbol{x}$

$$\boldsymbol{x} = [x_i] \quad i = 1, 2, ..., n \tag{2.55}$$

are governed by an initial value problem of the form[4]

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= f(\boldsymbol{x}, \boldsymbol{\theta}, t) \\ \boldsymbol{x}(0) &= \boldsymbol{x}_0(\boldsymbol{\theta}) \end{aligned} \tag{2.56}$$

$\boldsymbol{h}()$ is called the observer model because it relates how the observed variables in $\boldsymbol{y}$ depend upon the states and parameters. $\boldsymbol{\varepsilon}(t)$ in (2.54) represents model error as in previous sections.

Below, the notation is simplified slightly by replacing $\boldsymbol{y}(t)$ with $y(t)$ and $\boldsymbol{h}()$ with $h()$. In this thesis, we have only made use of this method for the case of a single observed variable (CO_2 concentration).

The problem is to solve the least squares problem

$$\min_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}) = \sum_{t=1}^{n_{obs}} [y(t) - h(t, \boldsymbol{x}, \boldsymbol{\theta})]^2 \tag{2.57}$$

subject to the initial value problem in (2.56).

In practice, the initial value problem could be replaced with a more general differential algebraic system. The basic challenge is the same: the observer is a nonlinear function of the parameters $\boldsymbol{\theta}$. Most methods for optimizing (2.57) require the gradient of the objective function. This can be difficult to obtain. Each component is

[4]The model can involve differential algebraic equations. We used the sensitivity method outlined here to estimate the parameters for the dynamic leaf model of Liu et al. [42], in which leaf area is governed by an ordinary differential equation. We introduce the method in the context of ODEs to keep the notation simple.

determined by

$$\begin{aligned} q_j &= \frac{\partial \phi}{\partial \theta_j} \\ &= -2 \sum_{t=1}^{n_{obs}} e_t \left(\frac{\partial h_t}{\partial \theta_j} + \frac{\partial h_t}{\partial \boldsymbol{x}_t} \frac{\partial \boldsymbol{x}_t}{\partial \theta_j} \right) \end{aligned} \tag{2.58}$$

where

$$e_t = y_t - h(t, \boldsymbol{x}, \boldsymbol{\theta}) \tag{2.59}$$

If a closed form solution exists for the initial value problem, then (2.58) can be evaluated analytically. If a solution does not exist, then one might consider using a finite difference approach to determine q_j. In general, however, sensitivity methods offer a more computationally efficient alternative. Sensitivity methods solve for sensitivity coefficients

$$S_{i,j} = \frac{\partial x_i}{\partial \theta_j} \tag{2.60}$$

It is convenient to collect these in the $n \times p$ sensitivity matrix as

$$\boldsymbol{S} = [S_{i,j}] \tag{2.61}$$

The sensitivity coefficients can be inserted into (2.58) to obtain the desired gradient.

$$q_j = -2 \sum_{t=1}^{n_{obs}} e_t \left(\frac{\partial h_t}{\partial \theta_j} + \frac{\partial h_t}{\partial \boldsymbol{x}_t} \boldsymbol{S}_{\cdot,j}(t) \right) \tag{2.62}$$

where $\boldsymbol{S}_{\cdot,j}$ is the j^{th} column of $\boldsymbol{S}$

Sensitivity methods fall into two categories.

- Forward methods. The sensitivity equations are solved forward from $t = 0$ to $t = T$ to get $\boldsymbol{S}(T)$. The sensivity equations are obtained by differentiating (2.56) with respect to $\boldsymbol{\theta}$. Using the chain rule,

$$\begin{aligned} \frac{d\boldsymbol{S}}{dt} &= \frac{\partial h}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial h}{\partial \boldsymbol{\theta}} \\ \boldsymbol{S}(0) &= \frac{\partial \boldsymbol{x}_0}{\partial \boldsymbol{\theta}} \end{aligned} \tag{2.63}$$

- Reverse (or adjoint) methods. The adjoint equations associated with the initial value problem are solved backward from $t = T$ to $t = 0$ to get $\boldsymbol{S}(t)$.

In this thesis, we use the forward sensitivity approach to obtain $\boldsymbol{S}(T)$. This matrix is used to calculate the gradient vector of the objective as in (2.62). A trust-region reflective method has been used to minimize the objective function in (2.57). The solution, $\hat{\boldsymbol{\theta}}$, is a local minimum of the objective.[5] The minimum that is obtained

[5]We anticipate a high degree of nonlinearity may be associated with the objective function,

implies a trajectory for $\boldsymbol{x}(t)$, and thus for the model predictions $y(\boldsymbol{x}, \hat{\boldsymbol{\theta}}, t)$. The variance of the model errors ($\boldsymbol{\varepsilon}(t)$) can be estimated using the sample variance of the residuals ($\hat{\sigma}$) determined in the usual manner. This value and the values of the sensitivity coefficients and partial derivatives of $h()$ at $\hat{\boldsymbol{\theta}}$ can be used to obtain the standard deviations of the least squares parameter estimates. These are obtained as the square roots of the diagonal terms in the covariance matrix

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \left(\boldsymbol{H}_{\boldsymbol{\theta}}^T \boldsymbol{H}_{\boldsymbol{\theta}}\right)^{-1} \hat{\sigma}^2 \tag{2.64}$$

where $\boldsymbol{H}_{\boldsymbol{\theta}}$ is the $n_{obs} \times p$ Jacobian matrix composed of the time dependent gradient vectors that showed up earlier in equation (2.62):

$$\boldsymbol{H}_{\boldsymbol{\theta}} = \begin{bmatrix} \cdots & \frac{\partial h_1}{\partial \theta_j} + \frac{\partial h_1}{\partial \boldsymbol{x}} \boldsymbol{S}_{\cdot,j}(t_1) & \cdots \\ \cdots & \frac{\partial h_2}{\partial \theta_j} + \frac{\partial h_2}{\partial \boldsymbol{x}} \boldsymbol{S}_{\cdot,j}(t_2) & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & \frac{\partial h_{n_{obs}}}{\partial \theta_j} + \frac{\partial h_{n_{obs}}}{\partial \boldsymbol{x}} \boldsymbol{S}_{\cdot,j}(t_{n_{obs}}) & \cdots \end{bmatrix} \tag{2.65}$$

2.10.5 Monte Carlo integration

Monte Carlo integration offers a sample based approach to (2.40) and (2.41). The idea is to use samples $\theta^{(k)}$ such that $k = 1, 2, ..., n$ to evaluate the expectation in (2.41). These samples are drawn from the posterior density. Thus, the solution to (2.40) is in the form of a set of samples and not a closed form density function. If a closed form density function were desired, one could use a density estimator to infer the parameters of a density function from the samples. The arithmetic average in (2.41) can be obtained from the samples by

$$\bar{f} = \frac{\sum_{k=1}^{n} \pi_k f(\theta^{(k)})}{\sum_{k=1}^{n} \pi_k} \tag{2.66}$$

The weights in (2.66) are not specified. In Monte Carlo integration, the samples are drawn according to the density $\pi(\theta|y)$ so the weightings are $\pi_k = 1/n$. The desired expectation is then approximately

$$\bar{f} = \frac{1}{n} \sum_{k=1}^{n} f(\theta^{(k)}) \tag{2.67}$$

The Metropolis-Hastings algorithm provides a general framework for sampling from $\pi(\theta|y)$. For example, the Metropolis algorithm, independence sampler and Gibbs sampler are all special cases of Metropolis-Hastings [35]. Given a sample $\theta^{(i)}$, the Metropolis-Hastings algorithm generates a candidate sample θ_c from a probing distribution $q(\theta_c|\theta^{(i)})$. The candidate sample is then accepted with probability

leading to multiple local minima. We have not used a technique simulated annealing to determine if the estimates we obtain are global minima. In practice, this would be desirable to ensure the solution being used for detrending is the most optimal one.

$$\alpha = min\left[1, \frac{\pi(\theta_c|y)q(\theta^{(i)}|\theta_c)}{\pi(\theta^{(i)}|y)q(\theta_c|\theta^{(i)})}\right] \tag{2.68}$$

If the candidate is rejected, then the present sample is kept. Thus, the update rule is

$$\theta^{(i+1)} = \begin{cases} \theta_c & \text{with probability } \alpha, \\ \theta^{(i)} & \text{otherwise.} \end{cases} \tag{2.69}$$

What differentiates the Metropolis algorithm, independence sampler, and Gibbs sampler from one another is the forms of their probing distributions $q(\theta_c|\theta^{(i)})$. Several factors affect choice of the probing distribution. Ideally, the probing density should be easy to compute, easy to sample from, and should have a shape similar to the posterior density $\pi(\theta|y)$.

2.10.6 Example: First-order reaction

Here, a simple example is introduced to demonstrate the effect of nonlinearity on the shape of the posterior distribution of least squares estimates. The problem is this. A series of observations are made of concentrations of a reactant in a first-order reaction. The reaction is A→B, with unknown reaction rate constant k and unknown initial concentration y_0. The model is thus

$$\begin{aligned} y(t) &= g(t, k, y_0) \\ &= y_0 \exp -kt \end{aligned} \tag{2.70}$$

Ten noisy observations of $y(t)$ have been recorded at different times t. The observations are summarized in Table 2.2. The question is, what is the associated

Table 2.2: Reactant concentrations for A→B

t (s)	y (mol L^{-1})
0	4.19
1	3.88
2	2.47
3	0.58
4	0.37
5	0.62
6	0.83
7	0.92
8	0.71
9	1.83
10	0.68

with the least squares estimates? To address this question, we did a couple things.

First, a steepest descent algorithm obtained the least squares estimates $(k, y_0) = (0.3165, 4.3175)$ in 32 iterations. The least squares predictions are shown in Figure 2-12.

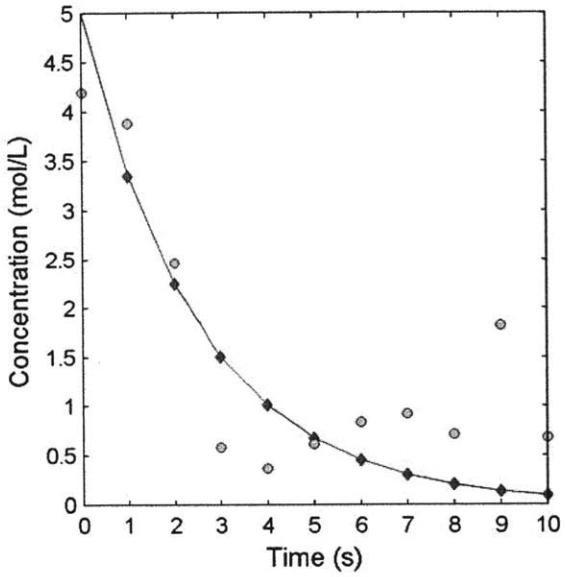

Figure 2-12: Reaction example: observed versus predicted. The observations are shown in blue circles. The first-order model predictions based on the least squares estimates $(k, y_0) = (0.3165, 4.3175)$ are shown in red diamonds.

The value of the Jacobian matrix $\boldsymbol{J} = [\partial g_t/\partial k, \partial g_t/\partial y_0]$ at the solution was determined to be

$$\bar{\boldsymbol{J}} = \begin{bmatrix} -3.146 & 0.729 \\ -4.585 & 0.531 \\ -5.012 & 0.387 \\ -4.869 & 0.282 \\ -4.435 & 0.206 \\ -3.878 & 0.150 \\ 0.109 & -3.297 \\ -2.746 & 0.080 \\ -2.251 & 0.058 \\ -1.823 & 0.042 \end{bmatrix}$$

Plugging this Jacobian into (2.49) gives us ellipsoidal confidence regions. We calculated the asymptotic confidence region and the exact confidence region given by (2.51) at the 0.05 and 0.01 levels of significance. We also generated Monte Carlo samples from the posterior distribution of (k, y_0).

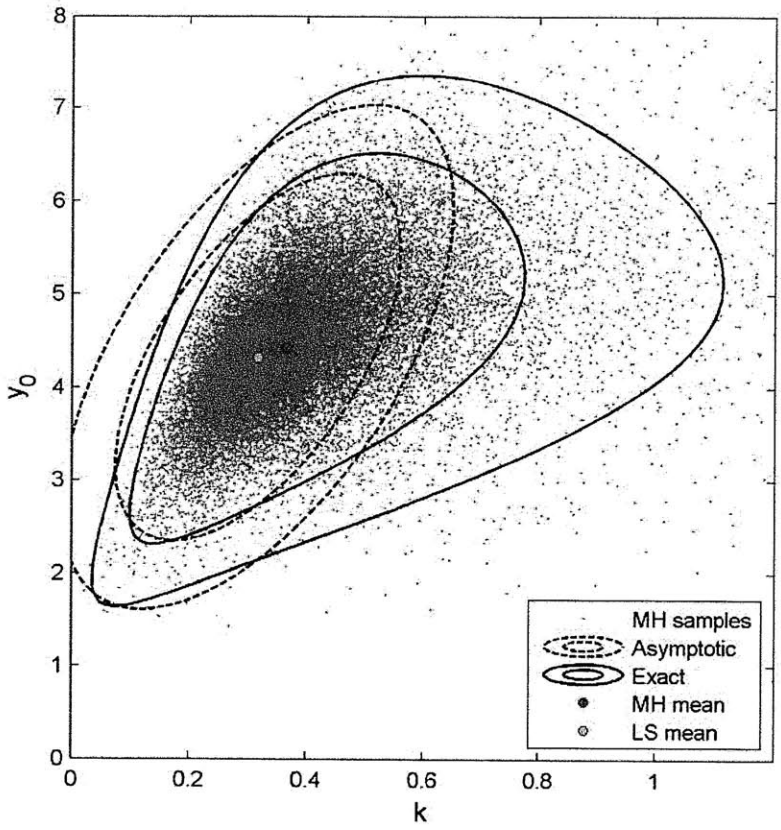

Figure 2-13: Reaction example: effect of nonlinearity on parametric uncertainty. The dashed contours are asymptotic confidence regions. The solid contours are exact confidence regions. For each type of contour, two isolines are shown. The inner isoline is that associated with an 0.05 level of significance. The outer isoline is based on an 0.01 level of significance. The grey dots represent 10^5 post-burn-in samples from the posterior density of (k, y_0) obtained using the Metropolis-Hastings algorithm. Also shown are the mean values of the parameters from least squares and Metropolis-Hastings. The least squares (LS, blue) and Bayesian (MH, red) estimates are also plotted.

From Figure 2-13, we observe that the asymptotic confidence regions poorly represent the actual posterior distribution of the parameters. The consequence is that the covariance matrix of the least squares estimates obtained using (2.50) alone cannot completely capture the uncertainty in the parameters. In such case the use of actual samples from the Bayesian posterior density may be more appropriate means of accounting for parametric uncertainty.

2.11 Autocorrelation

Most approaches to shift detection assume the model errors are independent. This is usually not the case. Many have studied the effects of autocorrelation and heteroskedasticity on inferences about the parameters [43, 44, 45, 46]. The consensus is that when residuals are autocorrelated, estimates of Δy obtained by standard estimation methods like ordinary least squares will tend to be overconfident[6] [43]. The implication for leak detection is that not accounting for autocorrelation in the residuals can lead to high false alarm rates, since naturally occurring shifts will tend to be detected more frequently than they should be. If residuals are found to be correlated, estimation methods exist capable of accounting for it should be used.

2.11.1 Autocorrelation

An error process $\varepsilon(t)$ is said to be correlated, autocorrelated, or serially correlated, at lag τ if

$$\rho(\tau) = \frac{\mathrm{E}\left[\varepsilon(t)\varepsilon(t-\tau)\right]}{\sigma(t)\sigma(t-\tau)} \neq 0 \tag{2.71}$$

for some lag time $\tau \neq 0$. Here, $\rho(\tau)$ is the correlation coefficient of the process at lag τ and $\sigma(t)$ is the standard deviation of the process at time t. $\rho(\tau)$ will be between -1 and 1. If the process is zero mean then the correlation at zero lag is

$$\begin{aligned} \rho(0) &= \frac{\mathrm{E}\left[u^2(t)\right]}{\sigma^2(t)} \\ &= \mathrm{Var}\left\{\varepsilon(t)\right\} / \mathrm{Var}\left\{\varepsilon(t)\right\} \\ &= 1 \end{aligned} \tag{2.72}$$

A process is said to be positively correlated if $\rho(\tau) > 0$, negatively correlated if $\rho(\tau) < 0$, and uncorrelated if $\rho(\tau) \equiv 0$.

When residuals are autocorrelated, the observations cannot be explained simply as the sum of model predictions and random errors. Two alternatives exist for modeling correlated observations. The first replaces the random error term ($\varepsilon(t)$) with an autoregressive error model. The second replaces it with a moving average error model. In the order p autoregressive model (often denoted AR(p)), the observations are given by

$$\begin{aligned} y(t_j) &= \mathcal{M}(\boldsymbol{\theta}, t_j) + \varepsilon(t_j) \\ \varepsilon(t_j) &= \sum_{k=1}^{p} \rho_k \varepsilon(t_{j-k}) + \xi(t) \end{aligned} \tag{2.73}$$

where $\varepsilon(t)$ is still a random error but now the model errors are $\varepsilon(t_j)$ which depend on past errors $\varepsilon(t_{j-k})$ through autoregressive coefficients ρ_k. The same order moving

[6] In the sense that uncertainty in Δy will tend to be underestimated, leading to confidence intervals that are too narrow.

average error model (denoted MA(p)) implies different behavior,

$$\begin{aligned} y(t_j) &= \mathcal{M}(\boldsymbol{\theta}, t_j) + \varepsilon(t_j) \\ \varepsilon(t_j) &= \sum_{k=1}^{p} \phi_k \varepsilon(t_{j-k}) + \xi(t) \end{aligned} \tag{2.74}$$

where $\xi(t_{j-k})$ are past random shocks or innovations. The difference between the AR(p) and MA(p) models is AR(p) remembers the magnitudes of past deviations $\varepsilon(t_{j-k})$, whereas MA(p) only remembers the magnitudes of historical shocks $\xi(t_{j-k})$. The implication is that Perturbations tend to persist longer in autoregressive processes than in moving average ones. In the leak detection problem, moving average processes are better suited to modeling disturbances due to factors like correlation due to the measurement procedure, whereas autoregressive are better for modeling correlation due to atmospheric mixing. Not all contributions to correlation have such physical grounds. An important contributor in practice might be due to missing and misrepresented physics in the detrending model for the observations. In general, it is unclear whether an AR or MA process is better. Therefore, both are usually considered together in the form of autoregressive moving average (ARMA) models.

2.11.2 Correlation coefficients

Since only a single realization $\boldsymbol{e} = [e_1, ..., e_N]^T$ of the process $\varepsilon(t)$ is available for leak detection[7], the autocorrelation coefficient for lag k can be estimated using the following version of the sample correlation coefficient:

$$\hat{\rho}(k) = \frac{(\boldsymbol{e}_{1:N-k} - \bar{e}_{1:N-k})^T (\boldsymbol{e}_{k+1:N} - \bar{e}_{k+1:N})}{(n-k) \cdot s_{1:N-k} s_{k+1:N}} \tag{2.75}$$

where $\boldsymbol{e}_{i:j} = [e_i, ..., e_j]^T$, and where $\bar{e}_{i:j}$ and $s_{i:j}$ are the sample mean and sample variance of $\boldsymbol{e}_{i:j}$, respectively, given by

$$\bar{e}_{i:j} = \frac{1}{j-i} \sum_{t=i}^{j} e(t) \tag{2.76}$$

$$s_{i:j} = \frac{1}{j-i-1} \sum_{t=i}^{j} [e(t) - \bar{e}_{i:j}]^2 \tag{2.77}$$

These equations assume samples are recorded at equal intervals and that there are no missing values in the series. When missing values are present, as is often the case for in situ CO_2 measurements, the equations can be modified to accomodate for these missing values as follows. First, form $\boldsymbol{e}_{1:N-k}$ and $\boldsymbol{e}_{k+1:N}$ as usual. Each vector is of length $N-k$. Some of the values in each vector are missing. Determine the $N(k)$ indices $\ell \in \{1, ..., N-k\}$ for which neither $\boldsymbol{e}_{1:N-k}(\ell)$ nor $\boldsymbol{e}_{k+1:N}(\ell)$ are missing, and

[7]No two sensors can measure CO_2 at the same point and time.

form new $N(k)$-length vectors $\tilde{\boldsymbol{u}}_1$ and $\tilde{\boldsymbol{u}}_2$ using these indices so that $\boldsymbol{e}_{1:N-k} \rightarrow \tilde{\boldsymbol{e}}_1$ and $\boldsymbol{e}_{k+1:N} \rightarrow \tilde{\boldsymbol{u}}_2$. Note that $N(k) \leq N-k$, depending on the locations of missing values in the original series. The sample correlation coefficient can then be calculated as

$$\hat{\rho}(k) = \frac{(\tilde{\boldsymbol{e}}_1 - \bar{e}_1)^T (\tilde{\boldsymbol{e}}_2 - \bar{e}_2)}{N(k) \cdot s_1 s_2} \tag{2.78}$$

where $\bar{e}_i$ and s_i ($i \in \{1,2\}$) are the sample mean and sample variance of $\tilde{\boldsymbol{e}}_i$, respectively, and are given by

$$\bar{e}_i = \frac{1}{N(k)} \sum_{t=i}^{N(k)} \tilde{e}_i(t) \tag{2.79}$$

$$s_i = \frac{1}{N(k)-1} \sum_{t=i}^{N(k)} [\tilde{e}_i(t) - \bar{e}_i]^2 \tag{2.80}$$

In practice, sample correlation coefficients will always differ from zero in a random fashion, even for truly white noise series. The question is, how to decide whether a series is autocorrelated given this element of chance?

2.11.3 Testing for autocorrelation

Many tests exist for autocorrelation [44, 47]. The data are plagued by gaps. An exploratory approach capable of handling these gaps involves constructing the Lomb periodogram (see Appendix B) of the residuals. We make use of this approach in later chapters to test models' abilities to remove systematic trends in the observations. Here, we describe two alternative approaches based on autocorrelation. The first is a Lagrange multiplier test [48, 49] proposed by Breusch and Godfrey in the late 1970's. The second is based on confidence intervals for sample correlation coefficients.

The Breusch-Godfrey test

Given a model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$, the Breusch-Godfrey test can be used to test for ARMA(p,q) behavior. The Breusch-Godfrey test for AR(p) behavior in the residuals proceeds as follows [50].

1. Detrend the observations to obtain $\boldsymbol{e}$. Estimate the variance (σ) of the model errors using the sample variance ($\hat{\sigma}$) of the residuals.

2. Use the residuals to construct the $N \times p$ lag matrix $\boldsymbol{E}_p = [\boldsymbol{e}_1 ... \boldsymbol{e}_p]$, where $\boldsymbol{e}_i = [0...0, e_1 ... e_{N-i}]^T$.

3. Calculate the test statistic

$$l = \boldsymbol{e}^T \boldsymbol{E}_p \left[\boldsymbol{E}_p^T \boldsymbol{E}_p - \boldsymbol{E}_p^T \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{E}_p\right]^{-1} \boldsymbol{E}_p^T \boldsymbol{e} / \hat{\sigma} \tag{2.81}$$

The null distribution of l is χ^2_p. Thus, if l exceeds a critical value of the χ^2_p distribution, then the null hypothesis that the $p \times 1$ AR coefficient vector $\boldsymbol{\rho} = 0$ is rejected in favor of the alternative that $\boldsymbol{\rho} \neq 0$.

Testing using confidence intervals

The second approach is an exploratory one wherein the 95% confidence intervals for $\hat{\rho}(k)$ as a function of lag k is plotted alongside the sample autocorrelation function $\hat{\rho}(k)$. If the samples in $\boldsymbol{\varepsilon}$ are independent and normally distributed, then the value obtained for $\hat{\rho}(k)$ will tend to lie in the interval

$$-\frac{1.96}{\sqrt{N(k)}} \leq \hat{\rho}(k) \leq +\frac{1.96}{\sqrt{n(k)}} \tag{2.82}$$

The value of 1.96 arises as the upper critical value of the standard normal distribution for which the probability of exceeding that value is $(1 - 95\%)/2$. If the sample autocorrelation function falls outside the interval defined by (2.82) at low lags (e.g., $k = 1, 2, ...$), then the hypothesis that the series is uncorrelated is rejected at the 0.05 level of significance.

Summary: testing for autocorrelation

The presence of $N(k)$ in the denominator of (2.82) indicates the effect gaps have on our ability to detect autocorrelation in a set of samples. Bigger gaps mean fewer samples to work with, implying wider confidence intervals and therefore decreased ability to decipher whether samples are autocorrelated. Gaps have a second, more serious consequence however and this is standard methods like Breusch-Godfrey and the confidence interval approach introduced here do not apply. These methods assume the samples are evenly spaced in time. Gaps will generally violate this assumption. In such case, alternative methods like the Lomb periodogram represent better alternatives because they are capable of detecting harmonic content in unevenly sampled signals.

2.11.4 Dealing with autocorrelation

Autocorrelation introduces off-diagonal terms into the error covariance matrix, so the correct assumption is in fact $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Omega})$. There are two ways to deal with autocorrelation. The first is to estimate the coefficients in $\boldsymbol{\Omega}$ along with the other unknowns in the error model. The downside of this approach is that this matrix is $N \times N$, so the number of unknowns grow like the square of the number of observations. The second approach, which is more common, is to truncate the error model after some number of terms. **Akaike's FPE** [51] is an example of a criterion for choosing the order of an autoregressive model (see 3.2.2). This can be combined with **Yule-Walker algorithm** [52] for estimating the regression coefficients (see 3.2.4).

2.12 Lessons for leak detection

Standard estimation procedures such as ordinary least squares will tend to lead to higher rates of false alarm when model errors are correlated over time. If systematic effects can be explained in $\mathcal{M}(\boldsymbol{\theta}, t)$ it may be possible to remove much of this autocorrelation by improving the detrending model. In general, some degree of autocorrelation will remain due to missing and misrepresented physics in the model and due to nonzero instrumental response. Therefore, when detecting leaks it will be important to use some method capable of accounting for autocorrelation.

Chapter 3

Illustrating statistical issues for leak detection

The leak detection problem requires considering shifts beyond historical levels in future observations of atmospheric CO_2. Doing so requires making predictions - more specifically, forecasts - about the values the observations will take in the near future and then waiting to see whether the values that are eventually observed coincide with our predictions. Were instruments perfect and nature predictable, detecting shifts would be a simple task. The dilemma is that instruments are always imprecise and nature unpredictable, so that uncertainty must be entertained. This section introduces a simple framework for thinking about the effects various sources of uncertainty in the leak detection problem, introduces important concepts affecting our ability to detect, and finishes with a real example illustrating the effect of autocorrelation on baseline estimation.

3.1 Illustrating the basic issues

Chapter 2 introduced a number of algorithms for detection. The algorithms were presented in the context of "N" observations. But how big should N be? This section introduces factors influencing our choice of N, and clarifies the impact this choice will have on the ability to detect in the context of hypothetical observations.

3.1.1 Hypothetical trajectories

Atmospheric CO_2 concentrations at Mauna Loa are presently around 390 ppm [5]. The scale of uncertainty associated associated with hourly averages at Mauna Loa is on the order of 0.2 ppm (one standard deviation, Komhyr et al. [53]). For simplicity, let us ignore calibration biases, assume that CO_2 concentrations remain at 390 ppm in the absence of leakage, and assume measurements are uncorrelated. A series of CO_2 measurements would then be given by

$$y(t) = a_0 + \varepsilon(t) \tag{3.1}$$

where $a_0 = 390$ ppm is the signal mean, and $\varepsilon(t)$ are random measurement errors with zero mean and standard deviation $a_1 = 0.2$ ppm.

In the classical case, the errors are Gaussian so $\varepsilon(t) \sim N(0, \alpha_1^2)$. The next subsection presents examples of trajectories obtained when the errors have different distributions, but the same mean and variance. These trajectories enable us to explore effects of non-normality. A further complication is to account for forward uncertainty. This is done to capture the forecasting aspect of the problem: the leak detection problem requires comparison of observed values with forecasted values which per Section 2.7.1 are associated with both observation and forward uncertainty. The scenarios considered in the sections to follow are enumerated in Table 3.1

Table 3.1: Hypothetical trajectories for shift detection at Mauna Loa

Scenario	Description	Distribution	Variance (ppm^2)
G	Classical, neglecting forecast	Gaussian	0.2^2
Gf	Classical, with forecast good to ± 1 ppm (2σ)	Gaussian	$0.2^2 + 1^2$
U	Uniform, neglecting forecast	Uniform	0.2^2
L	Lognormal, neglecting forecast	Lognormal	0.2^2
B	Bimodal, neglecting forecast	Bimodal	0.2^2

[a] Trajectories are defined in terms of error distribution and variance. Mean was set to 390 ppm in all cases. The number of observations (N) is sometimes varied, so it will be defined later as needed.

Figure 3-1 illustrates trajectories and probability density functions based on the Gaussian and non-Gaussian scenarios defined in Table 3.1.

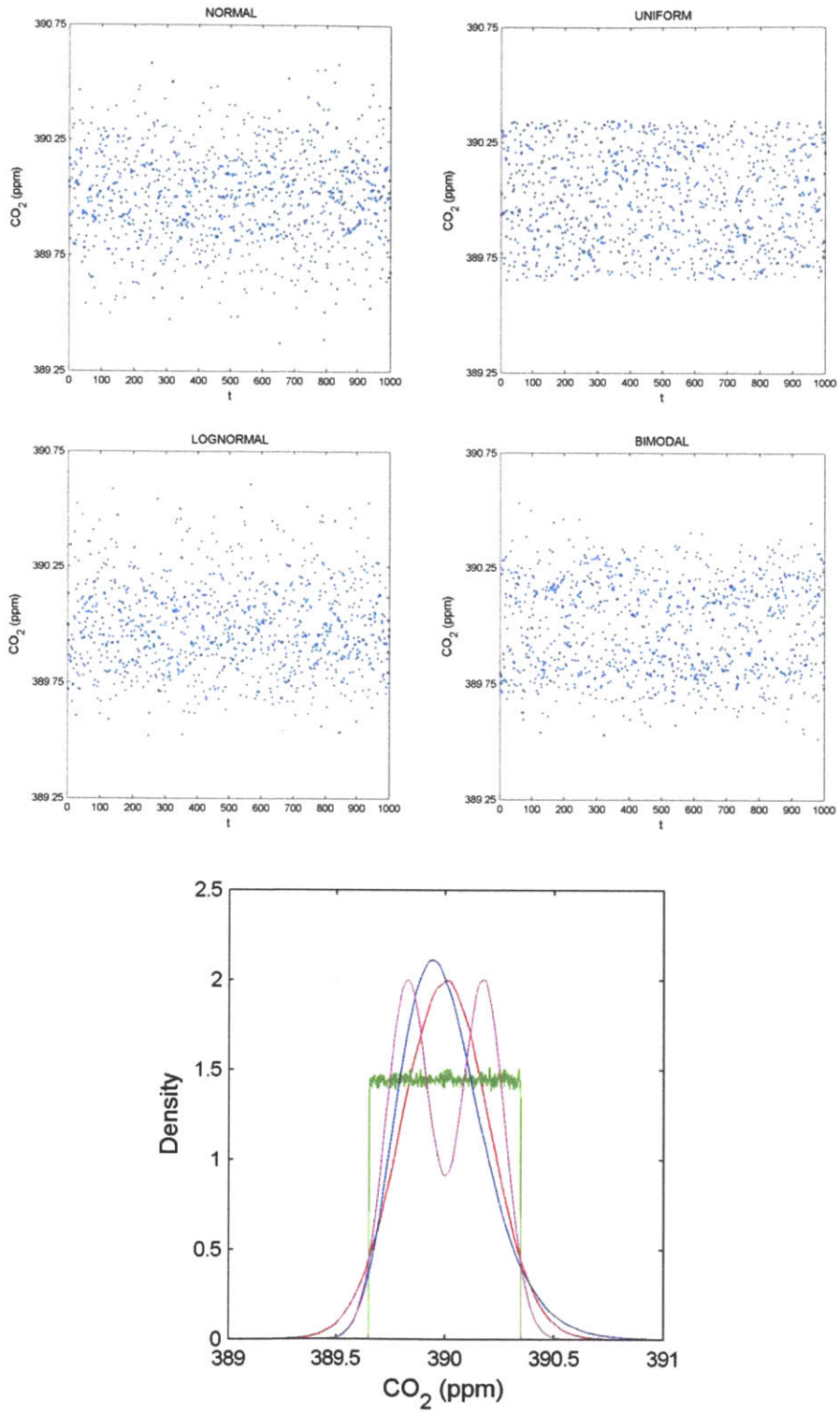

Figure 3-1: Four trajectories with equal mean and variance but different underlying distributions. Top two rows: trajectories of 10^3 values having the same mean and standard deviation (390 ppm and 10 ppm, respectively) but coming from very different distributions. From top left to bottom right: (1) normal, (2) uniform, (3) lognormal, and (4) bimodal distributions. Bottom: kernel density estimates based on 10^6 draws from each of the four distributions.

3.1.2 One-tailed Z-test

Detecting a mean shift requires an estimate of the mean before the shift arrives. For simplicity, we assume the mean (a_0) and variance (a_1^2) of the observations are both known exactly and that the latter remains constant in time. The previous chapter showed how Z-tests can be applied to detect shifts in CO_2 data. It was shown that by increasing the signal to noise ratio and/or amount of data considered the false positive rate associated with detection could be reduced. The example in this section assumes a leak is absent to begin with, and uses a right-tailed Z-test to test for upward shifts linked to leakage. A right-tailed Z-test with $\alpha = 0.05$ results in a critical value of 1.64. This is the value z for which $\Pr(\mathcal{T} \geq z|\mathcal{L}_0) = 0.05$. That is, α is the probability of detecting a shift when none is present. This is the **false alarm rate** associated with the test. The level of significance α is therefore the false alarm rate the user is willing to put up with in search of actual shifts, and should be chosen as such. Controlling α is important because it means having control over the amount of effort spent looking in more detail for leaks that are not present.

What if a leak *is* present? If the sensor is sensitive to the leak's plume in the atmosphere, then leakage will induce a positive shift of magnitude Δy in the observations. On average, (2.8) will be

$$\mathcal{T} = (\Delta y/a_1)\sqrt{N} \tag{3.2}$$

The signal to noise ratio of a random signal is given by the inverse of the coefficient of variation, which is the ratio of standard deviation to mean. Therefore, $\Delta y/a_1$ is just the **signal to noise ratio (SNR)** of the shift:

$$SNR = \Delta y/a_1 \tag{3.3}$$

The distribution of the test statistic in the presence of a shift is $\mathcal{T} \sim N(SNR\sqrt{N}, 1)$. For $\alpha = 0.05$, the **detection rate**[1] is

$$\begin{aligned} P_d &= \Pr(\mathcal{T} \geq 1.64|\mathcal{L}_1) \\ &= 1 - F_z(1.64 - SNR\sqrt{N}) \end{aligned} \tag{3.4}$$

where F_z is the cumulative probability density function for the standard normal distribution.

According to (3.4), the probability of detecting a shift increases with the signal to noise ratio of the shift (SNR) and with the number of samples observed (N). Thus, our chance of detecting a shift improves as:

- The magnitude of the shift (Δy) increases.
- The variance of the forecast (a_1^2) decreases.
- The number of observations (N) increases.

[1] Also referred to as the probability of detection.

3.1.3 Evolution of uncertainty in the sample mean

The Z-test assumes the mean of the observations (a_0) is known. When it is unknown to begin with, it must be estimated by a sample mean. The question is, how many samples (N) are required to estimate the mean? The 95% confidence interval of the sample mean is given by

$$a_0 \pm 1.96 \cdot a_1/\sqrt{N} \tag{3.5}$$

The confidence interval narrows with increasing sample size N as illustrated in Figure 3-2. For a test like the Z-test, this means the baseline estimate improves as N increases. Since the detection problem requires the comparison of two such sample means, increasing N should also decrease the shift size that is detectable. Therefore, the narrower the confidence interval, the smaller the shift we expect to be able to detect.

3.1.4 Limits to detection

Figure 3-2 demonstrates the effects of increasing sample size (N) and increasing shift size (Δy) on our ability to detect a shift in a white Gaussian noise with mean of 390 ppm and standard deviation of 10 ppm. As expected, increasing the number of samples considered decreases the shift size that can be detected. [2]

Without increasing the number of sensors or measurement frequency, increasing N means waiting longer to detect a shift. Therefore, Figure 3-2 illustrates the tradeoff between **time to detection** and detectable shift size.

[2]The latter two figures have been determined by calculating the probabilities of detection at the 0.05 level of significance for varying sample size and by increasing the number of samples available until the detection probability exceeds 90%.

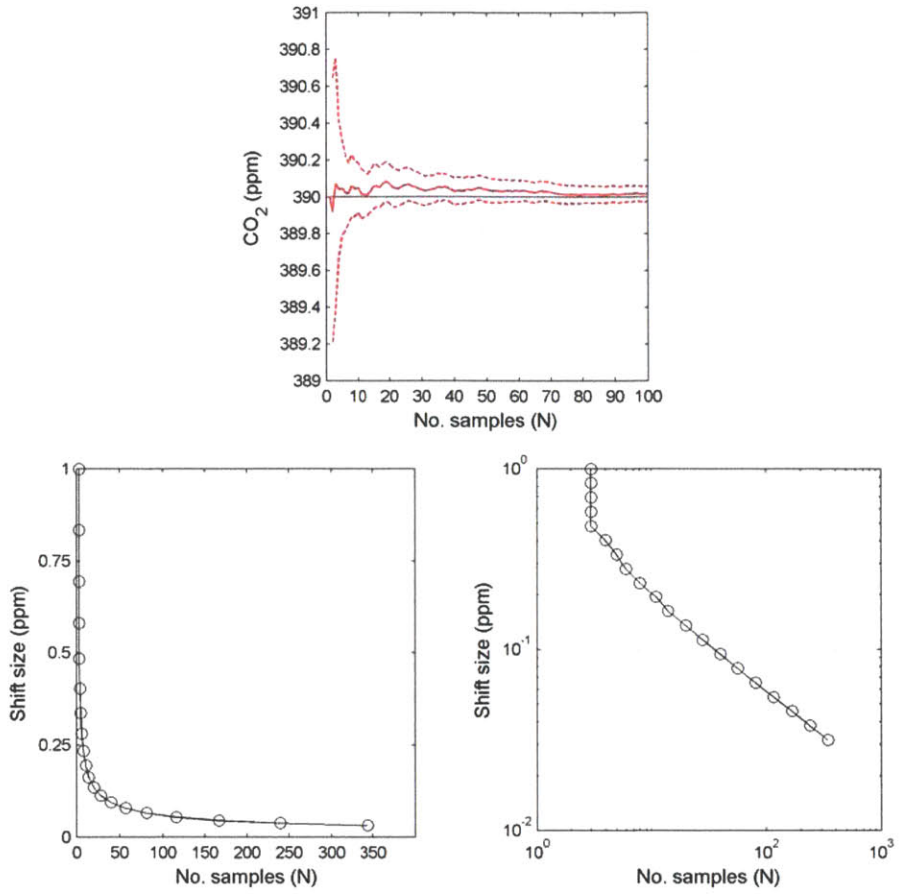

Figure 3-2: Evolution of sample statistics. Top: the red solid line shows evolution of sample mean as more samples are considered. The dashed red lines show the 95% confidence intervals computed as $\bar{y} \pm 1.96 \cdot a_1/\sqrt{N}$. Bottom: the effect of increasing sample size on the size of the shift that is detectable at the 0.05 level of significance. Here, time to detection was calculated as the number of samples required to achieve 90% detection rate at the 0.05 level of significance, using (3.4) to determine probability of detection. The figure on the right is the same as the one on the left, but plotted as a log-log plot to emphasize the logarithmic dependence between detection time and shift size. **An important consequence of this relationship is that there are diminishing returns from waiting longer.**

3.1.5 Normality and the Central Limit theorem

The results of the Z-test depend on critical values which assume the sample mean is normally distributed. If the errors $a_1\varepsilon(t)$ are Guassian, then the sample mean will also be Gaussian. What happens if instead of being Gaussian the measurement errors are uniform or lognormal? Figure 3-3 below considers the density of the sample mean of uniform random values as the number of samples considered grows. As the number of samples increases, the sample mean appears to become increasingly more Gaussian. This is in accord with the Central Limit theorem, which states that the sum of an increasingly large number of independent random variables tends to a normal distribution. The consequence is that if the samples are not Gaussian, then

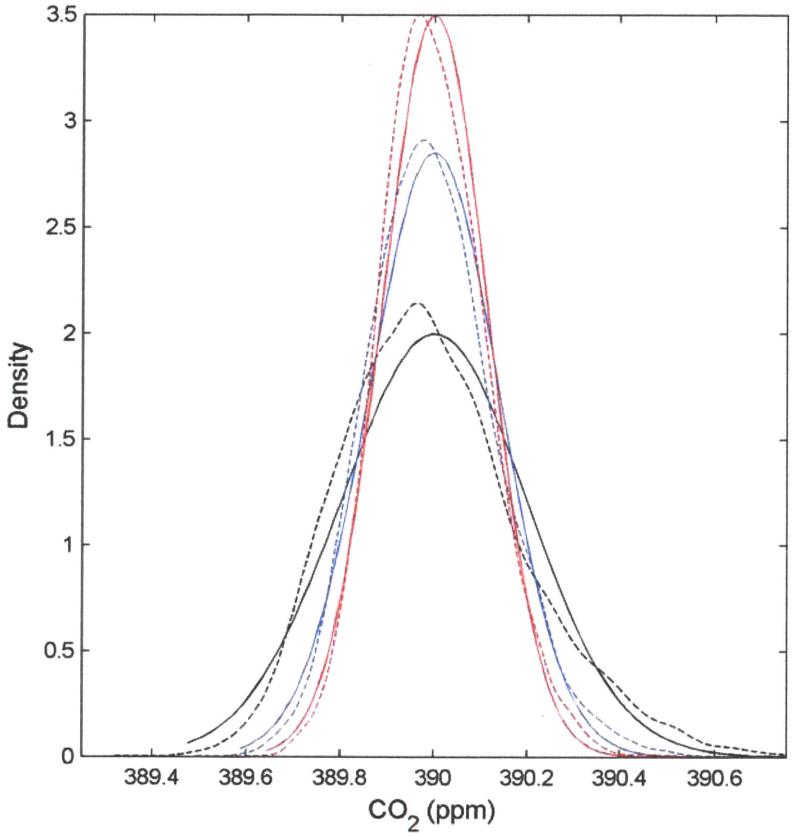

Figure 3-3: Normality and sample means. This figure depicts the Central Limit theorem in action. Each figure shows empirical fits (obtained using kernel density estimators, dashed) and normal density fits (solid) to 5×10^3 sample means generated by considering increasing numbers of observations. Black: $N = 1$ observation. Blue: $N = 10^2$. Bottom: $N = 5 \times 10^3$. The observations were generated from Scenario L in Table 3.1.

probabilities determined using Z-test statistics will differ somewhat from what they should be. Figure 3-4 uses the distributions introduced earlier in the context of Figure 3-1 to assess the importance of distribution type. According to the figure, the 95% confidence intervals differ most when only a few samples are considered, but they converge quickly to those assuming normality as sample size increases. Bayesian methods should be more used when means are being compared and sample size is limited.

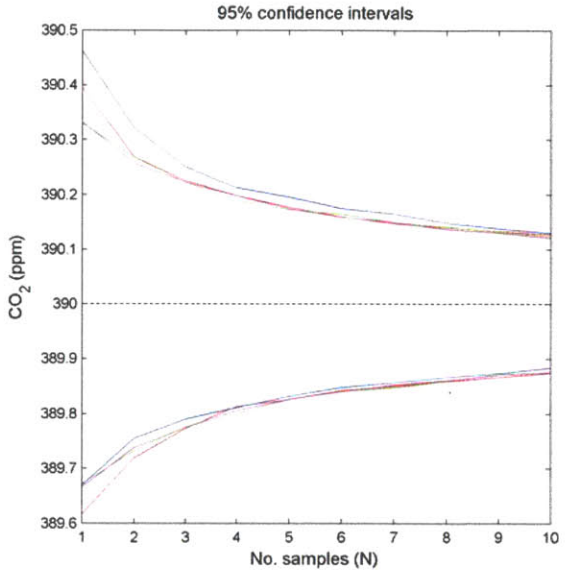

Figure 3-4: Comparing evolutions of sample mean confidence intervals for non normal and normal observations. The dashed black line is the mean used to generate the observations, which is 390 ppm. The curves surrounding the mean are 95% confidence intervals generated by simulating random trajectories from the various populations being compared and then computing standard errors from the standard deviations of these random trajectories. Red: normal observations, green: uniform, blue: lognormal, magenta: bimodal.

3.1.6 The effect of forward uncertainty

The uncertainty associated with the model parameters will also affect our ability to detect. The simplest possible case is where the mean predicted by the model is uncertain. The observations are then given by

$$y(t) = a_0 + a_0'(\omega) + \varepsilon(t) \tag{3.6}$$

$a_0'(\omega)$ is a random number that is independent of time and which represents error of the modeled mean from the actual mean of the signal. Scenario Nf in Table 3.1 has been constructed assuming Gaussian errors with zero mean and variances $\mathrm{Var}\{a_0'\} = 2.5^2$ and $\mathrm{Var}\{\varepsilon(t)\} = a_1$. We can consider generating random trajectories based on scenarios N and Nf. In case Nf, the trajectories will consist of random values having mean a_0 and standard deviation $\sqrt{(a_0')^2 + a_1^2}$. Therefore, uncertainty in the

model parameters inflates the noise associated with the forecasted observations. This reduces the signal to noise ratio from $SNR = \Delta y/0.2$ to $SNR = \Delta y/\sqrt{0.2^2 + 2.5^2}$. Such a change implies reduced detection rates and worse **limits to detection**[3]. The time to detection for a given shift size should also increase. Figure 3-5 demonstrates the impact of forward uncertainty on time to detection.

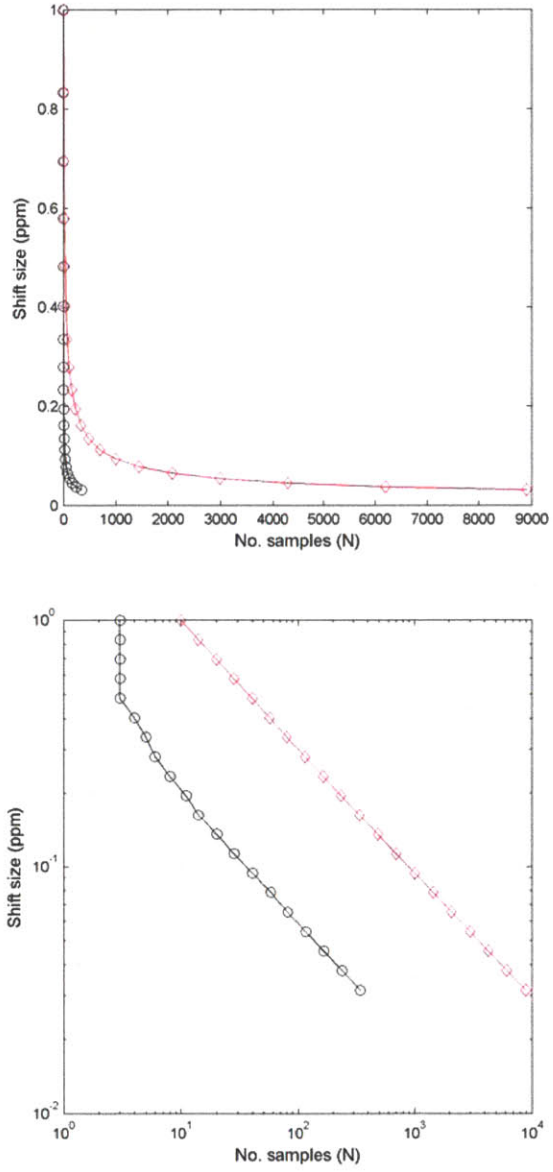

Figure 3-5: Time to detection and parametric uncertainty. Shown are number of samples required to achieve 90% probability of detection for varied shift size. Two cases from Table 3.1 are considered, N and Nf.

[3]We use "limits to detection" to refer to curves trading-off detection time T against shift size Δy of the type presented in Section 3.1.4

3.1.7 Summary

There is a tradeoff between the size shift (Δy) that can be detected and the number of samples (N) required to detect. Without adding sensors or increasing sampling frequency, the only way to decrease the detectable size shift is to wait longer. For monitoring and verification of leakage from geological sequestration, waiting longer may not be an option: it means increased risk to the population and increased loss of CO_2 before mitigation. The basic analysis presented in this section suggests that smaller leaks can be detected in another way, which is to increase the signal to noise ratio of the shift. Practically, this means reducing uncertainty associated with forecasts. This can be achieved in two ways: (1) increasing measurement precision or (2) improving the model so the variance of the residuals decreases.

The number of samples considered also has implications for the false alarm rate. There are two reasons for this. First, the more samples considered the more accurate the sample means being compared will be. Second, when classical detection algorithms like Z-tests and t-tests are being used but the model errors are not normally distributed, increasing sample size tends to increase the validity of the Central Limit theorem. In general, when the errors are not Gaussian it should be safer to use a Bayesian detection algorithm since the number of samples required for the Central Limit theorem to apply is not the same for all distributions.

The above results all assume the samples are independent. Autocorrelation is also an important issue for mean shift detection. The next section illustrates the consequences of and importance of using estimation methods capable of handling autocorrelated errors.

3.2 Autocorrelation and the case for improved models

This section (1) illustrates that modeling reduces autocorrelation and improves the validity of the normality hypothesis and (2) demonstrates the danger of not accounting for autocorrelation in the residuals. The message is clear: explain the observations as best as possible before testing for a shift. These ideas are generated in the context of atmospheric CO_2 data from Mauna Loa. The data, which comprise monthly averages of daily mean values from NOAA (post-1974) and SIO (pre-1974) and were obtained from `ftp://ftp.cmdl.noaa.gov/ccg/co2/trends/co2_mm_mlo.txt`, are plotted in Figure 3-6.

Three models are used in this section:

- Model Z. The model is

$$y(t) = \beta_1 + \varepsilon(t) \tag{3.7}$$

  where time t is measured in days, $y(t)$ is monthly average CO_2 at Mauna Loa, and β_1 is a model parameter representing the mean concentration during the training period. The model, which is a zero-slope straight line, characterizes the ability to predict CO_2 in the absence of any prior knowledge of the trends

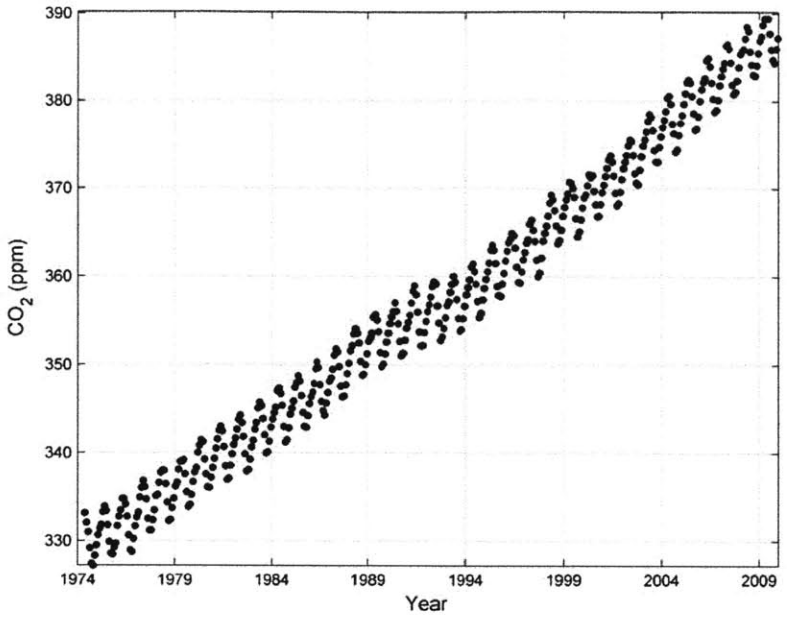

Figure 3-6: Monthly mean CO_2 data from Mauna Loa.

present in atmospheric CO_2. The properties of the residuals will be identical to the properties of the original data.

- Model S. The model is

$$y(t) = \beta_1 + \beta_2(t - t_0) + \varepsilon(t) \tag{3.8}$$

where t, $y(t)$, and β_1 are the same as in Z and β_2 is a model parameter representing the mean growth rate of CO_2 concentration during the training period. This model is a slope-intercept model capable of capturing the increasing trend in CO_2, assuming a constant growth rate in time.

- Model SR'. The model is

$$y(t) = \beta_1 + \beta_2(t - t_0) + \beta_3 SOLRAD(latitude, t - \beta_4') + \varepsilon(t) \tag{3.9}$$

where t, $y(t)$, β_1, and β_2 are the same as in S. $SOLRAD(latitude, t)$ is the flux of solar radiation reaching the Earth's outer atmosphere at latitude *latitude* at solar noon of the day in which time t falls. The solar radiation model is described in Chapter 5. It requires two parameters. β_3 represents the yield of CO_2 uptake by photosynthesis per unit of solar radiation. β_4 represents the lag time required for biological growth as sunlight starts to increase in the spring. The work done in this section modified the model by setting β_4 manually. The value chosen was obtained by adjusting β_4 and using the value that minimized

the RSS. The "optimal" value for Mauna Loa was 1.2 months.

3.2.1 Impact of detrending on autocorrelation, normality

The impact of detrending on autocorrelation and normality was explored using exploratory plots and significance tests. Since the detrending models are all linear, of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} \tag{3.10}$$

fits were obtained using the ordinary least squares (OLS) estimates of the parameters ($\hat{\boldsymbol{\beta}}$). These were obtained as follows. First, OLS estimates of the regression coefficients $\boldsymbol{\beta}$ were computed considering all available observations. The resulting OLS estimates, $\hat{\boldsymbol{\beta}}$, were combined with the explanatory variables to generate OLS predictions according to $\boldsymbol{X}\hat{\boldsymbol{\beta}}$. The difference between the observations and the OLS predictions yielded OLS residuals $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The residuals were then tested for autocorrelation in two ways. The first was by plotting the autocorrelation function and comparing the result to the 95% confidence interval associated with the hypothesis that the residuals be independent and normally distributed. The second way used the Breusch-Godfrey test from Section 2.11.3 against the alternative that the residuals were first-order autoregressive. The residuals were tested for normality using three significance tests. These were the chi-square goodness-of-fit test, Lilliefors test, and the Jarque-Bera test from Section 2.8.3.

The Breusch-Godfrey test rejected the hypothesis that the residuals were independent at the 0.05 level of significance for all models considered. The test statistic decreased (from 423.7 to 319.4 to 293.0) as model complexity increased, hinting that the level of autocorrelation present was decreasing with the explanatory power of the model. This observation was corroborated by the autocorrelation function plots. In these plots, it was observed that the sample autocorrelation function of the residuals vanishes to zero more rapidly as the detrending model improves. These results suggest that **by removing more of what can be explained in atmospheric CO_2 records, the effects of autocorrelation should be reduced.** Note that the effects of autocorrelation have not yet been demonstrated. It has simply been shown that autocorrelation is present in the residuals, and that the level of autocorrelation present appears to decrease as the detrending model improves. The next section deals with the consequences of autocorrelation for the risk involved in detecting mean shifts in noisy signals.

Similar conclusions applied to the normality hypothesis. Normality was rejected by all three tests for model Z, by two of the three tests for model S, and by none of the tests for model SR' at the 0.05 level of significance. Essentially, the residuals become more normal as the model is improved. This said, however, the only model for which classical detection methods are justified is model SR'. For the other models, Bayesian methods are justified but classical methods might be robust if the normality assumption is not overly poor. The impact of using least squares instead of Bayesian estimation was not assessed. Assuming the least squares esimates used to obtain the residuals we tested are not significantly biased from their Bayesian counterparts, our

results seem to suggest that improving the detrending model leads to more Gaussian residuals.

3.2.2 Choosing the order of an autoregressive model (Akaike's FPE)

Akaike [51] introduced a method for determining the order of an autoregressive moving average (ARMA) model that values parsimony and therefore avoids overfitting. We used this method to test for autoregressive (AR) behavior. The underlying model is

$$\begin{aligned} \varepsilon_t &= \rho_0 + \sum_{i=1}^{q} \rho_i \varepsilon_{t-i} + \xi_t \\ \xi_t &\sim N\left(0, \sigma^2\right) \end{aligned} \tag{3.11}$$

where $\varepsilon(t)$ is the model error and $\xi(t)$ is a random shock. Given residuals $e(t)$, the questions is, are the model errors autoregressive? For N observations, there will be an equal number of residuals, $\{e_t : t = 1, 2, ..., N\}$. The algorithm analyzes the residuals for AR behavior as follows:

1. Remove the sample mean from the residuals.

$$\tilde{e}_t = e_t - \bar{e} \tag{3.12}$$

   where

$$\bar{e} = \frac{1}{N} \sum_{t=1}^{N} e_t \tag{3.13}$$

2. Set an upper bound q_{hi} on the model order to be considered. q_{hi} should be large enough to include the minimum q_0 sought in step 4, but should also generally be much smaller than the number of lags necessary to estimate the power spectrum of $e(t)$ by Fourier transforming the windowed sample autocovariance function.

3. Choose a model order $q = 1$. Calculate the sample autocovariances

$$\hat{\gamma}(h) = \frac{1}{N} \sum_{t=1}^{N-h} \tilde{e}_t \tilde{e}_{t+h} \quad h = 0, 1, ..., q_{hi} \tag{3.14}$$

4. Determine the values $\{\rho_i : i = 1, 2, ..., q\}$ that minimize the sum of square residuals (RSS)

$$RSS^{(q)} = RSS\left[\rho^{(\mathbf{q})}\right] = \sum_{t=1}^{N} \left(\tilde{e}_t - \sum_{i=1}^{q} \rho_i^{(q)} \tilde{e}_{t-i} \right)^2 \tag{3.15}$$

   The sum of square residuals is calculated assuming

$$\tilde{e}_0 = \tilde{e}_{-1} = \cdots = \tilde{e}_{1-q} = 0 \tag{3.16}$$

This is equivalent to solving the Yule-Walker equation

$$\hat{\mathbf{\Gamma}}_{\mathbf{q}}\hat{\rho}^{(\mathbf{q})} = \hat{\gamma}_{\mathbf{q}} \Leftrightarrow \hat{\rho}^{(\mathbf{q})} = \hat{\mathbf{\Gamma}}_{\mathbf{q}}^{-\mathbf{1}}\hat{\gamma}_{\mathbf{q}} \tag{3.17}$$

The variance of the random errors is estimated by

$$S^{(q)} = \frac{1}{N-1-q} RSS^{(q)} \tag{3.18}$$

The '1' in the denominator comes from the reduction of the number of degrees of freedom due to the presence of ρ_0. When ρ_0 is absent, which occurs when the u_t's are residuals from a detrending model involving a constant term β_0, then

$$S^{(q)} = \frac{1}{N-p-q} RSS^{(q)} \tag{3.19}$$

where p is the number of systematic model parameters. In the special case that the observations are already zero mean, so that both systematic model parameters and ρ_0 are absent from the model,

$$S^{(q)} = \frac{1}{N-q} RSS^{(q)} \tag{3.20}$$

Taking the first case as an example, Akaike defines the final prediction error (FPE) associated with the autoregressive model of order q as

$$FPE^{(q)} = \left(1 + \frac{q+1}{N}\right) S^{(q)} \tag{3.21}$$

5. Choose the value q_0 of q that minimizes the final prediction error $FPE(q)$, where q is of course restricted to integers equal to and between 0 and q_{hi}.

The algorithm is clarified through examples. Here, we include the figure used to diagnose AR model order for the Mauna Loa example in Chapter 3.2.

Akaike noted that FPE tends to be large when q is either too large or too small. The reason this happens when q is too small is a consequence of missing terms in the autoregressive model. I believe the reason it happens when q is too large is related to the assumption in (3.16), but it can also be anticipated because (3.21) is a growing function of q.

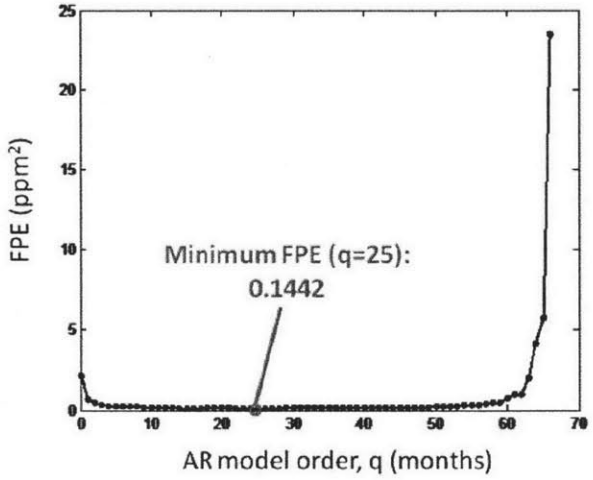

Figure 3-7: Akaike's FPE plot for the monthly Mauna Loa OLS residuals from model SR'.

3.2.3 Dealing with autocorrelation

OLS assumes the residuals are independent and normally distributed. When residuals are autocorrelated, generalized least squares (GLS) is justified. Unlike GLS, OLS precludes the error covariance matrix from containing anything but zeros off-diagonal. But what consequences does this have for the ability to detect a shift in the mean of a set of residuals?

To illustrate the effect of autocorrelation on our ability to detect a shift, we fitted SR' to the monthly CO_2 data in Figure 3-6 using OLS and estimated GLS (EGLS). The results are described in Section 3.2.5. Model SR' was selected because its OLS residuals were previously found to be autocorrelated but normally distributed, so normality was not an issue. OLS fitting was done in the usual way. EGLS was performed using the **Yule-Walker algorithm** [52] (see 3.2.4), with autoregressive model order (of 25 months) determined by minimizing **Akaike's FPE** [51] (see 3.2.2).

3.2.4 The Yule-Walker Algorithm

Autocorrelation introduces off-diagonal terms into the error covariance matrix, so the correct assumption is in fact $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Omega})$. There are two ways to deal with the autocorrelation. The first is to estimate the coefficients in $\boldsymbol{\Omega}$ along with the other unknowns in the error model. The downside of this approach is that this matrix is $N \times N$, so the number of unknowns grow like the square of the number of observations. The second approach, which is more common, is to truncate the error model after some number of terms. **Akaike's FPE** [51] is an example of a criterion for choosing the order of an autoregressive model (see Appendix 3.2.2). This can be combined with **Yule-Walker algorithm** [52] for determining the regression coefficients. The latter algorithm, which is sometimes referred to as "estimated generalized leasts squares" (EGLS), is presented below.

Gallant and Goebel [52] introduced a method of estimating model parameters when the residuals are autocorrelated. It is assumes the order of the disturbance (error) model - which can be either of the autoregressive or moving average types - is already specified. The model is assumed to take the form

$$\begin{aligned} y(t) &= \mathcal{M}(\boldsymbol{\theta}) + \varepsilon(t) \\ \varepsilon(t) &= -\sum_{i=1}^{q} \rho_i \varepsilon(t - i\Delta t) + \xi(t) \\ \xi(t) &\sim N\left(0, \sigma^2\right) \end{aligned} \tag{3.22}$$

where $\boldsymbol{\mathcal{M}}(\boldsymbol{\theta}) = [\mathcal{M}(\boldsymbol{\theta}, t_i)]$. The errors $\xi(t)$ are assumed to be independent, zero mean, and homoskedastic with variance σ^2. Gallant and Goebel make the additional assumption that the population from which the $\varepsilon(t)$'s are drawn is covariance stationary. The implication of this assumption is that the variance-covariance matrix of the model errors $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)^T$ has a special banded structure.

The algorithm employed by Gallant and Goebel proceeds as follows:

1. Compute the ordinary nonlinear least squares estimator $\hat{\boldsymbol{\theta}}$. The obtained value minimizes

$$[\boldsymbol{y} - \boldsymbol{\mathcal{M}}(\boldsymbol{\theta})]^T [\boldsymbol{y} - \boldsymbol{\mathcal{M}}(\boldsymbol{\theta})] \tag{3.23}$$

The variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\boldsymbol{V}} = \hat{\sigma}^2 \hat{\boldsymbol{C}} \tag{3.24}$$

where

$$\hat{\sigma}^2 = \frac{1}{N-p}\left[\boldsymbol{y} - \boldsymbol{\mathcal{M}}\left(\hat{\boldsymbol{\theta}}\right)\right]^T \left[\boldsymbol{y} - \boldsymbol{\mathcal{M}}\left(\hat{\boldsymbol{\theta}}\right)\right] \tag{3.25}$$

estimates the variance of the disturbances (which ordinary least squares assumes to be uncorrelated) and

$$\hat{\boldsymbol{C}} = \left[\boldsymbol{J}^T(\hat{\boldsymbol{\theta}}) \cdot \boldsymbol{J}(\hat{\boldsymbol{\theta}})\right]^{-1} \tag{3.26}$$

$\boldsymbol{J}$ is the $N \times p$ Jacobian of the model predictions ($\boldsymbol{\mathcal{M}}$) with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{J}(\boldsymbol{\theta}) = \begin{bmatrix} \nabla \mathcal{M}(t_1, \boldsymbol{\theta}) \\ \nabla \mathcal{M}(t_2, \boldsymbol{\theta}) \\ \vdots \\ \nabla \mathcal{M}(t_N, \boldsymbol{\theta}) \end{bmatrix} \tag{3.27}$$

2. Compute the residuals

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{\mathcal{M}}\left(\hat{\boldsymbol{\theta}}\right) \tag{3.28}$$

The sample autocovariances up to lag q are calculated as

$$\hat{\gamma}(h) = \frac{1}{N} \sum_{t=1}^{N-h} e_t e_{t+h\Delta t} \quad h = 0, 1, ..., q \quad q < N - p \tag{3.29}$$

3. Compute least squares estimates of the autoregressive coefficients and the error variance using the Yule-Walker equations:

$$\begin{aligned} \hat{\boldsymbol{\rho}} &= -\hat{\boldsymbol{\Gamma}}_q^{-1} \hat{\boldsymbol{\gamma}}_q \\ \hat{\sigma}^2 &= \hat{\gamma}(0) + \hat{\boldsymbol{\rho}}^T \hat{\boldsymbol{\gamma}}_q \end{aligned} \tag{3.30}$$

where

$$\hat{\boldsymbol{\rho}} = [\hat{\rho}(1), \hat{\rho}(2), ..., \hat{\rho}(q)]^T \tag{3.31}$$

and

$$\hat{\boldsymbol{\Gamma}}_q = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(q-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(q-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}(q-1) & \hat{\gamma}(q-2) & \cdots & \hat{\gamma}(0) \end{bmatrix} \tag{3.32}$$

where

$$\hat{\boldsymbol{\gamma}}_q = [\hat{\gamma}(1), \hat{\gamma}(2), ..., \hat{\gamma}(q)]^T \tag{3.33}$$

Next, factor $\hat{\boldsymbol{\Gamma}}_q^{-1} = \hat{\boldsymbol{P}}_q^T \hat{\boldsymbol{P}}_q$ using, for example, Cholesky decomposition. The following steps are described in terms of an nxn matrix $\hat{\boldsymbol{P}}$. For large n, this matrix can be expensive to store. The beauty of Gallant and Goebel's algorithm is that $\hat{\boldsymbol{P}}$ need not actually be stored. All the necessary calculations involving $\hat{\boldsymbol{P}}$ can be performed just by knowing (1) the qxq matrix $\hat{\boldsymbol{P}}_q$, (2) the q autoregressive coefficients $\hat{\boldsymbol{\rho}}$, and (3) the ordinary least squares variance $\hat{\sigma}^2$. Thus, rather than storing N^2 values we need only store $q^2 + q$. We leave the reader to refer to Gallant and Goebel's original paper [52] for the expression for $\hat{\boldsymbol{P}}$. Their notation differs slightly from ours. By replacing their $\hat{a}_i$ with our $\hat{\rho}_i$ one obtains the desired expression.

4. Compute the estimated generalized nonlinear least squares model parameter

estimates by minimizing

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N}\left[\hat{\boldsymbol{P}}\boldsymbol{y} - \boldsymbol{P}\hat{\boldsymbol{\mathcal{M}}}(\boldsymbol{\theta})\right]^T \left[\hat{\boldsymbol{P}}\boldsymbol{y} - \boldsymbol{P}\hat{\boldsymbol{\mathcal{M}}}(\boldsymbol{\theta})\right] \tag{3.34}$$

with respect to $\boldsymbol{\theta}$. The solution to this problem, which we denote $\tilde{\boldsymbol{\theta}}$, is a generalized least squares estimate which accounts for autocorrelation in the ordinary least squares residuals obtained in (3.28). From (3.34) we see that the generalized least squares estimate is obtained by transforming the systematic model by $\hat{\boldsymbol{P}}$ and then using ordinary least squares to estimate the model parameters from the transformed model. Transforming the model in the way described makes the residuals uncorrelated, which is why ordinary least squares is applicable. From $\tilde{\boldsymbol{\theta}}$ we obtain

$$\tilde{\sigma^2} = \frac{1}{N-p}\left[\hat{\boldsymbol{P}}\boldsymbol{y} - \boldsymbol{P}\hat{\boldsymbol{\mathcal{M}}}(\boldsymbol{\theta})\right]^T \left[\hat{\boldsymbol{P}}\boldsymbol{y} - \boldsymbol{P}\hat{\boldsymbol{\mathcal{M}}}(\boldsymbol{\theta})\right] \tag{3.35}$$

which estimates the variance of the random errors in (3.22), and

$$\tilde{\boldsymbol{C}} = \left[\boldsymbol{J}^T(\tilde{\boldsymbol{\theta}}) \cdot \hat{\boldsymbol{P}}^T\hat{\boldsymbol{P}} \cdot \boldsymbol{J}(\tilde{\boldsymbol{\theta}})\right]^{-1} \tag{3.36}$$

The variance-covariance matrix of the estimator is

$$\tilde{\boldsymbol{V}}_{\boldsymbol{\theta}} = \tilde{\sigma}^2\tilde{\boldsymbol{C}} \tag{3.37}$$

Standard errors can be estimated as usual from the diagonal terms of this matrix.

5. (Optional.) Set $\hat{\boldsymbol{\theta}}$ equal to $\tilde{\boldsymbol{\theta}}$ and repeat steps 2-4.

3.2.5 OLS and GLS results

The 95% standard errors associated with the OLS and EGLS estimates of the regression coefficients β_1, β_2, and β_3 were determined from the diagonal of the covariance matrix of the least squares estimates. These are provided along with the least squares estimates in Table 3.2.

Table 3.2: OLS and EGLS results for model SR'

	OLS results		EGLS results, AR(25)	
Parameter	Value	Standard error	Value	Standard error
β_1	345.95	0.66	346.84	1.60
β_2	4.45e-03	1.88e-05	4.45e-03	1.10e-04
β_3	-1.45e-02	5.26e-04	-1.49e-02	1.08e-03

When autocorrelation is considered (as demonstrated by the EGLS case), the standard errors associated with the model parameters are larger than they are when autocorrelation is neglected (as in the OLS case.) Although in general it is impossible to say whether autocorrelation will inflate or deflate uncertainty, the former appears to be the usual case. In their chapter on serial correlation, Davidson and MacKinnon [44]) demonstrate that positive AR(1) autoregressive behavior leads to inflated errors, and similar results tend to hold for various other extended cases ([43, 46, 54]).

3.2.6 Conclusions

Building better detrending models appears to reduce autocorrelation in the residuals and improve the normality assumption. Autocorrelation injects additional risk into the detection process because OLS leads us to be overconfident in the mean values being compared, which means higher rates of false alarm when detecting mean shifts. This section demonstrates the need for models capable of capturing systematic trends in the observations to reduce this added risk.

Chapter 4

Data sources and dataset characterization

Our ability to detect will depend upon the nature of the observations. The goal of this chapter is to analyze some existing data sets and characterize their properties.

4.1 Variability of CO_2 in the atmosphere

Historic CO_2 data exhibit many trends. There is an increasing interannual trend due to fossil fuel emissions. A similar interannual trend is present at all latitudes, but a latitude gradient of about 4 ppm exists with CO_2 levels being higher in the Northern Hemisphere. Small deviations from the fossil fuel based interannual trend occur that are less well understood. These have been hypothesized to arise because of temperature effects on terrestrial photosynthesis and respiration and on equilibrium partitioning of carbon dioxide across the air-sea interface. Most sites exhibit a seasonal cycle superimposed on the rising trend due to terrestrial photosynthesis and respiration. The magnitude of this cycle is much larger in the Northern Hemisphere, and ranges from 1-15 ppm. The phase and amplitude of the seasonal cycle have been observed to change systematically over time. This section reviews some of the trends that have been highlighted in the scientific literature.

4.1.1 $^{13}C/^{12}C$ ratios

Measuring carbon and oxygen isotopes in CO_2 is the most common way to determine where the CO_2 is coming from. Isotopes can distinguish the contributions of multiple sources to a mixture [55, 56, 57]. The ratio of ^{13}C to ^{12}C abundance in CO_2 can be used to partition changes in CO_2 into contributions from fossil fuel combustion, oceanic carbon, etc. About 1% of carbon exists as ^{13}C, with the remainder being mostly ^{12}C. Isotopic composition of carbon in CO_2 is typically defined relative to a reference value $r_{ref} \approx 0.01$ ‰ by

$$^{13}\delta = 1000 \times (r/r_{ref} - 1) \tag{4.1}$$

where r is the $^{13}C/^{12}C$ ratio of the air being measured. The $^{13}\delta$ of atmospheric CO_2 is about -8‰, plant carbon is around -25‰. Oceanic ^{13}C content is close to that of atmospheric CO_2 and fossil fuel, being composed mostly of fossilized plants, tends to have a ^{13}C content similar to plant carbon. These values are approximate. Actual values can be expected to vary with space, time, and other factors [58]. The $^{13}\delta$ value of a mixture of CO_2 from two sources is approximately given by sum of the mole-fraction weighted $^{13}\delta$'s of the original sources:

$$^{13}\delta_{mix} \approx \chi_1^{13}\delta_1 + \chi_2^{13}\delta_2 \tag{4.2}$$

where $\chi_i = N_i/(\sum N_i)$, N_i is the numer of moles from source 'i' in the mixture, and $^{13}\delta_{mix}$ is the $^{13}\delta$ of the mixture. Thus $^{13}\delta$ of CO_2 in the atmosphere will shift slightly from -8 towards -25‰ as the terrestrial contribution grows. Given knowledge of the fossil fuel contribution from elsewhere, $^{13}\delta$ can therefore be useful for constraining the individual contributions from oceanic and terrestrial sources.

4.1.2 Interannual trends

Long term atmospheric CO_2 observations are characterized by interannual trends. Similar rising trends are apparent at sites in both the Northern and Southern Hemispheres. This rising trend appears to be accelerating with increasing fossil fuel emissions. Small systematic anomalies are present in the emissions predicted trend that are harder to explain. The interannual trends are superimposed on a latitude gradient with CO_2 levels being higher in the Northern Hemisphere.

Atmospheric CO_2 is on the rise. Hofmann et al. [59] estimate CO_2 is doubling every 30 years with growth tracking population increases. The ice core records indicate that historic CO_2 mixing ratios ranged between 180 and 300 ppm [3]. Between 1750 and 2005, the global average CO_2 mixing ratio increased from somewhere between 275 and 285 ppm to 379 ppm, with half the change coming in the last ~30 years [4]. The global average CO_2 mixing ratio was approximately 386 ppm in 2009 and has been increasing at a rate of 0.66-2.92 ppm yr^{-1} since 1980 [5].

Increasing trends in CO_2 are predicted fairly well by cumulative emissions from fossil fuel combustion. Keeling et al. [60] propose a relationship between deseasonalized CO_2 and net fossil fuel combustion which assumes a constant fraction (f_{abf}) of emissions remain airborne. The model is

$$c_{inter}(t) = f_{abf} \int_{t_0}^{t} E(t)dt + c_{anom}(t) \tag{4.3}$$

where c_{inter} is the interannual trend neglecting seasonal and diurnal variations, $E(t)$ is the CO_2 emissions rate from fossil fuel combustion, and c_{anom} is the anomalous difference between the observed interannual trend and the trend predicted by fossil fuel emissions. Figure 4-1 compares cumulative emissions with ambient CO_2 from Mauna Loa. The emissions curve has been rescaled so it matches the concentration curve in Jan 1959 and Jan 1982. The emissions curve explains the observations quite closely. Systematic deviations of up to 2 ppm exist that cannot be explained by this

Figure 4-1: Mauna Loa CO_2 versus fossil fuel CO_2 emissions courtesy of Keeling et al. [60]. Shown are deseasonalized CO_2 obtained by removing a third order polynomial and harmonic signal from observations (dots) alongside rescaled CO_2 emissions (line). As described by the original authors, emissions have been rescaled to match the concentrations obtained in Jan 1959 and Jan 1982.

simple model. Reasons that have been given for these systematic departures include (1) El Nino events and (2) Indian summer monsoons. More recently, Francey et al. [61] accounts for volcanic emissions when correlating growth in CO_2 with fossil fuel emissions.

Keeling et al. [60] observed a growing interhemispheric gradient in CO_2 which correlates well with changes in fossil fuel emissions. The difference between deseasonalized CO_2 at Mauna Loa and the South Pole was estimated to have grown by about 2 ppm from 1958 to 1988. Growth slowed after the 1973 oil crisis, signaling the importance of fossil fuel emissions. The rationale is that most CO_2 emissions occur in the Northern Hemisphere and tend to accumulate there due to the presence of an equatorial barrier to atmospheric transport. Interhemispheric mixing occurs on time scales of 1-1.5 years. The interhemispheric gradient is not explained by fossil fuel emissions alone. If CO_2 is regressed against fossil fuel emissions and the resulting curve extrapolated back to zero cumulative emissions, Keeling et al. [60] find that the fit predicts higher CO_2 in the Southern Hemisphere than in the North. The reason for this has been the subject of some debate (see [62]).

4.1.3 Seasonal trends

Most sites have a seasonal cycle superimposed over the increasing trend described above. Past interpretations of seasonal cycles have been aided by isotopic data. At least three characteristics of seasonal trends have been studied. First, the seasonal amplitude is minimal in the Southern hemisphere and grows from South to North. Second, relative isotopic abundances are less negative at tropical latitudes. Third, the amplitude of the seasonal cycle has undergone systematic changes in time.

Keeling et al. [60] summarize characteristics of CO_2 and $^{13}\delta$ observations from numerous global observing sites. Isotopic data suggests the majority of seasonality is due to photosynthesis and respiration. During the summer, plants fix CO_2 by photosynthesis. This leads to lower CO_2 in the summer than in the winter. The

amplitude of the seasonal cycle is smaller in the Southern Hemisphere where there is less land mass for terrestrial uptake and release. The seasonal cycle also increases farther north of the equator. This is due to increasing seasonality in solar radiation, temperature and leaf mass. At Point Barrow, Alaska (~70° North), the seasonal cycle has an amplitude of 15-20 ppm. At Maua Loa, it is about 5 ppm. South of the equator, the seasonal cycle is typically at most 1 ppm.

Less negative $^{13}\delta$'s in the tropics indicate relatively small contributions from sources besides land plants to seasonal cycles there. Keeling et al. [60] hypothesize that the presence of plants with lower $^{13}\delta$'s and increased contributions from oceanic carbon might be behind these observations. There is evidence that seasonal amplitudes are increasing. Keeling et al. [10] observe 20-40% amplitude increases from 1964-1994 that are correlated with lengthened growing seasons. Buermann et al. [63] hypothesize the seasonal amplitude at Mauna Loa can be explained by alternating contributions from enhanced photosynthesis in North America and respiration in Eurasia.

4.1.4 Diurnal trends

Diurnal fluctuations are affected by a number of factors, including terrestrial photosynthesis and respiration that vary over the course of the day [64]. Fossil fuel combustion for heating and transportation can vary greatly with time of day. Boundary layer dynamics tend to exhibit diurnal behavior. Diurnal amplitudes are inflated when meteorology covaries with trends in surface sources. Changes in wind direction can drive variability when a site is located near the boundary separating spatially heterogeneous regions.

Law et al. [65] compare predictions of diurnal CO_2 variability in the atmospheric boundary layer obtained for a number of different locations using twenty five different models. Their study highlights the difficulty in predicting diurnal fluctuations. In a later study, Law et al. [66] focus on a single site at Cape Grimm, Tasmania. Their results support the importance of high resolution flux estimates in regions that are spatially heterogeneous. Cape Grimm is prone to urban contamination from Melbourne to the north and biomass burning in Tasmania. The authors improved the correlation between model predictions and the observations by ignoring records having large CO and H_2 concentrations, which are correlated with fossil fuel combustion, and selecting observations originating over the ocean to the west. Ahmadov et al. [67] demonstrate better predictions of diurnal trends by including subgridscale fluxes and meteorological transport. Their model consists of a nested scheme with a maximum resolution of 2 km trained using local meteorology and tracers. Flux measurements were used to constrain contributions from different types of plants. Simulations for the high resolution model (WRF-VPRM) are compared to the those from lower resolution models in Figure 4-2. The results indicate the higher resolution model is much better equipped to explain the timing and magnitudes of diurnal scale changes.

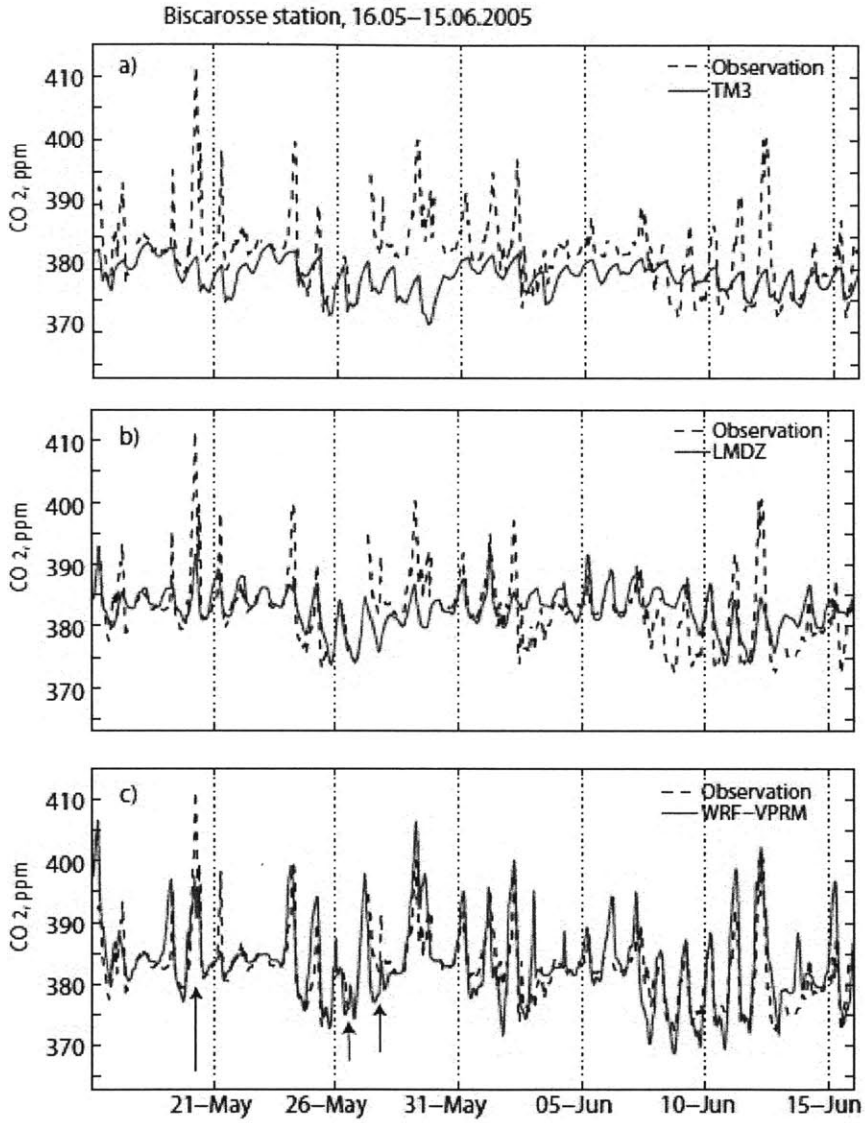

Figure 4-2: Effect of increasing model resolution on diurnal CO_2 predictions, adapted from Ahmadov et al. [67]. The top two figures show simulations (red) from low resolution models. The bottom figure shows the simulated trajectory from the 2 km resolution WRF-VPRM model. Observations are shown by dashed black lines. Arrows in the bottom figure indicate the presence of double peaks resulting from front passage or sea breeze WRF-VPRM can explain.

4.2 Data sources

This section characterizes atmospheric CO_2 records from three different sources: (1) Scripps and NOAA data from the Mauna Loa Observatory; (2) Harvard Forest; and (3) the Salt Lake CO_2 Network. These were chosen to represent a range of CO_2 emissions scenarios. The Mauna Loa records are unique because they are obtained from a high altitude location with barren surroundings and are mostly devoid of local influences. Harvard Forest is an ecological monitoring site in semi-rural, central Massachusetts. The Salt Lake CO_2 Network is an urban atmospheric monitoring network located in Salt Lake City, Utah.

There are numerous sources of CO_2 data and analyses; three key sites are:

- The **CarbonTracker**, **Interactive Atmospheric Data Visualization**, and **GLOBALVIEW** tools provided by NOAA at `http://www.esrl.noaa.gov/gmd/dv/`.

- CDIAC's **Online Trends** resource, which provides pointers to data and data visualization products and is available online at `http://cdiac.ornl.gov/trends/co2/contents.htm`.

- Scripps Insitution of Oceanography's CO_2 Program website, which can be found online at `http://scrippsco2.ucsd.edu/home/index.php`.

4.2.1 Mauna Loa

The Mauna Loa Observatory, located on the slope of Mauna Loa in Hawaii (see Figure 4-4), is unique because its surroundings are extremely barren and the site is located at ~3400 meters altitude so that the measurements come from above the boundary layer and are relatively free of local influences. Because of its unique location, the site's measurements are representative of global trends.

Data

There are two measurement programs at Mauna Loa. Scripps Institution of Oceanography started measuring in 1958. NOAA added continuous in situ measurements of their own in 1974 for comparison to the Scripps set up.

Both programs provide flask and in situ measurements of CO_2. Flask samples are typically collected twice [68]. In situ samples are made semi-continuously using infrared gas analyzers calibrated hourly. The measurements from the two programs (SIO and NOAA) come from distinct instruments.

Other variables measured at the Mauna Loa Observatory include $^{13}C/^{12}C$, $^{14}C/^{12}C$, and $^{18}O/^{16}O$ isotope ratios; atmospheric concentrations of hydrocarbons, halocarbons, CO, H_2, N_2O, O_3, and SF_6; and meteorology including wind speed and direction, barometric pressure, air temperature and precipitation.

Trends

Previous studies at Mauna Loa have made the following observations:

- *Increasing trend.* CO_2 has been increasing at Mauna Loa since Dave Keeling began measuring there in 1958. Figure 4-3 depicts annual increases in CO_2. The annual increase has been positive every year and appears to follow an increasing trend. The figure shows an exponential fit to the annual changes to show the trend the changes should be following if the data are growing exponentially like fossil fuel emissions [59].

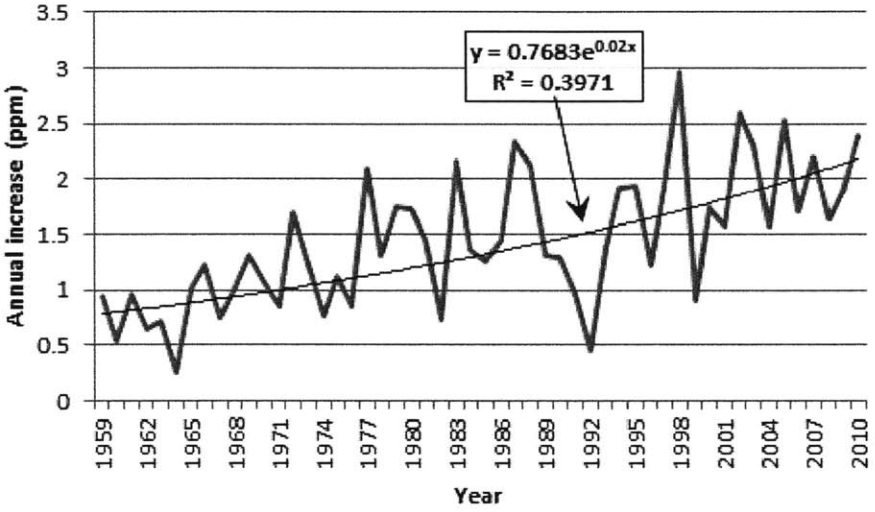

Figure 4-3: Annual increases in CO_2 at Mauna Loa, data courtesy of Dr. Pieter Tans, NOAA/ESRL, (`www.esrl.noaa.gov/gmd/ccgg/trends/`). Annual changes computed for Jan 1 to Dec 31 of each year are shown in blue. The black line is an exponential fit to the annual data, highlighting the increasing trend that appears to be present.

- *Seasonal trend.* Keeling et al. [69] determine the peak-to-peak seasonal amplitude to be around 5 ppm using an empirical fit composed of a cubic polynomial to represent the long term increase and 6- and 12-month harmonics to represent the seasonal cycle.

- *Changing seasonality.* The seasonal amplitude increased by 0.7% per year from 1958-1982 [70]. This translates to a 23% change in 30 years, which is consistent with Keeling et al.'s [10] estimate of 20-40% for 1964-1994. The size of the seasonal amplitude at a given site is affected by changes in upwind sources and sinks. Buermann et al. [63] determine that the growing seasonal amplitude at Mauna Loa is due to increased drawdown alternated with increased respiration, and that the amplitude stopped growing in the 1990's because drought reduced photosynthetic uptake in North America and because of changes in atmospheric circulation leading to less respiration enriched air from Eurasia.

- *Diurnal trends.* The measurements at Mauna Loa tend to exhibit little diurnal variability due to the location. The site is above the atmospheric boundary layer and surrounded by volcanic rock. Local influences cannot be avoided altogether, however. These are primarily a concern late at night, when radiative cooling brings air from upslope [71, 72, 73]. The air flowing downslope brings with it CO_2 emitted during outgassing events at the volcano's summit located 6-7 km away. Ryan [74] used CO_2 measurements to estimate emissions following three volcanic events. At night, emissions led to concentrations that were tens of ppm above background.

Figure 4-5 shows the data at three time scales. The interannual and seasonal trends already discussed are clearly present. Changes in seasonality and acceleration of the upward trend are difficult to see by eye. As expected, diurnal variability is fairly random.

4.2.2 Harvard Forest

Harvard Forest is one of the world's oldest and best equipped ecological monitoring sites. The semi-rural site is located in central Massachusetts approximately 100 miles west of Boston (see Figure 4-6). Prevailing wind directions are from the west-northwest and southwest. As a member of Fluxnet, one of the site's primary purposes is to provide measurements of CO_2 exchange between the local canopy and the atmosphere. Unlike most Fluxnet sites, however, the CO_2 concentrations used to generate eddy fluxes at Harvard Forest are calibrated frequently and are considered fairly accurate.

Data

CO_2 concentrations at Harvard Forest are measured in two ways. First, concentrations are measured above the canopy at ~30 meters. Second, concentrations are measured at several heights below the canopy. The first type of measurement is used in combination with wind velocity data to generate eddy fluxes. The second type provides a vertical profile which is used to measure CO_2 storage within the canopy. All measurements are made using infrared gas analyzers. A custom calibration procedure is applied hourly to maintain accuracy of the measurements.

Many other types of data are available besides CO_2. The EMS tower itself measures wind speed and direction, hourly precipitation, air temperature and pressure, and solar radiation. Field measurements of plant and soil biomass have been performed periodically at various locations. Past monitoring has also included measurements of atmospheric CO, NO, NO_2 and O_3 concentrations.

We obtained hourly mean CO_2 concentration data from the final quality-assured dataset made available at `http://atmos.seas.harvard.edu/lab/data/nigec-data.html`. The CO_2 data used in this thesis came from the above-canopy sampling location. The dataset also includes hourly wind direction and speed, air temperature, soil temperature, air pressure, relative humidity, precipitation, and CO, NO_y and O_3 concentrations.

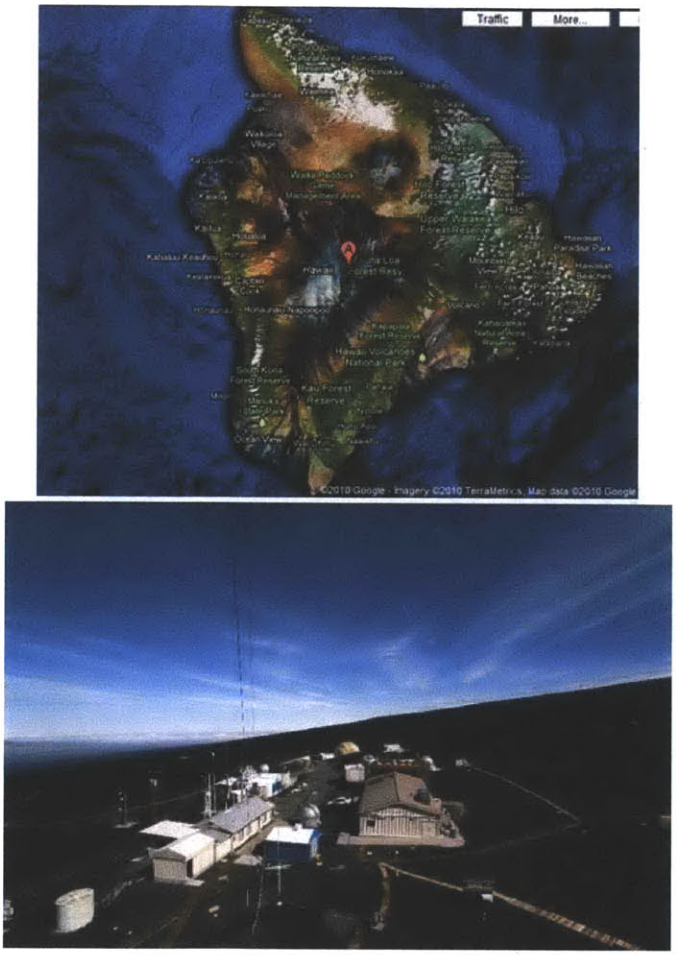

Figure 4-4: NOAA's monitoring site at Mauna Loa. Top: map of the island courtesy of Google Maps. Bottom: photograph of the Mauna Loa Observatory. The observatory is located at 19.5° north, 155.6° west, and 3,397 meters (11,142 feet) above sea level. The high altitude and the barren location make it an ideal location for measuring atmospheric concentrations above the boundary layer.

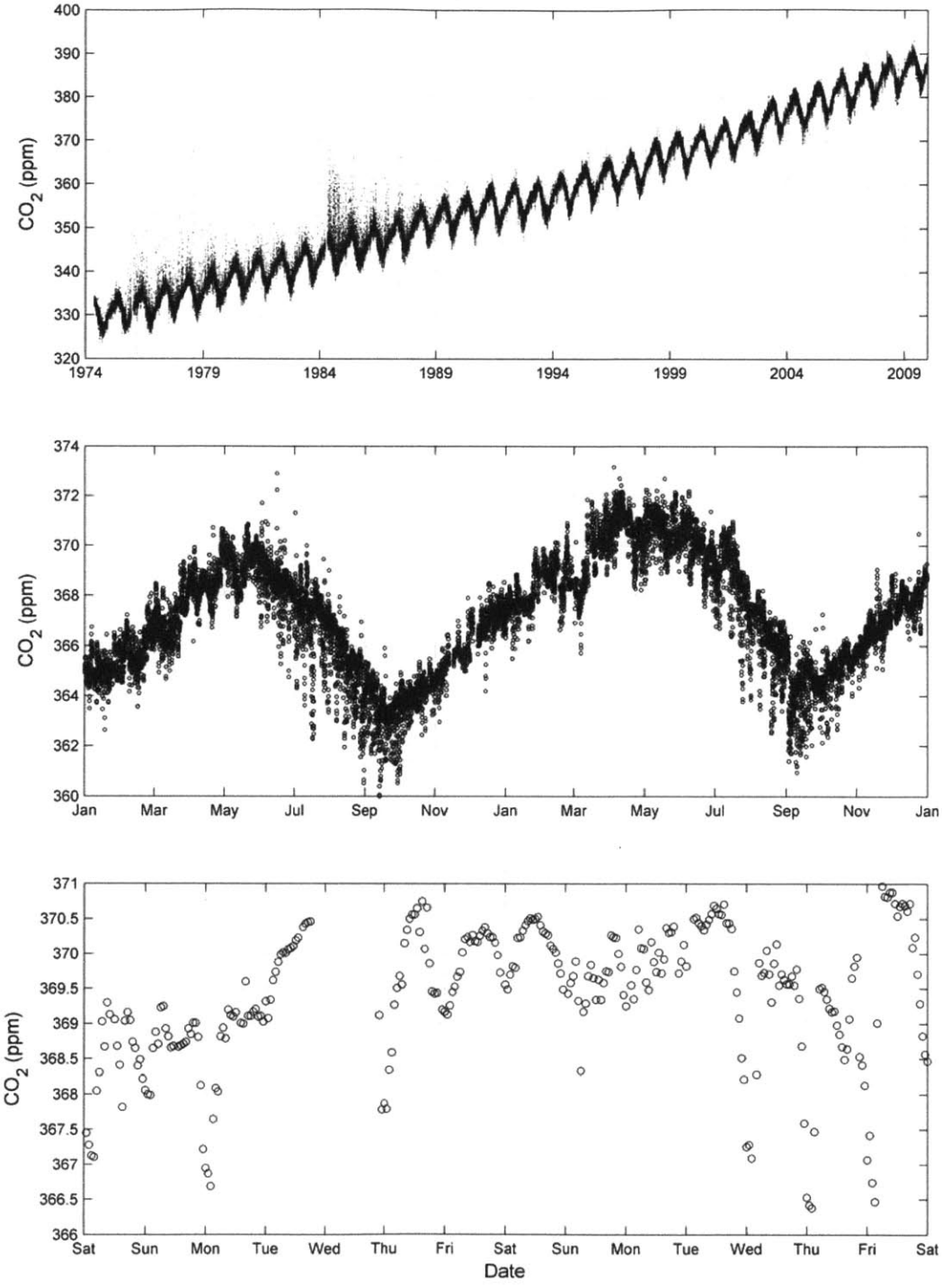

Figure 4-5: Atmospheric CO_2 trends at Mauna Loa. Top: complete hourly mean CO_2 time series from NOAA. Middle: zoom-in to observations between Jan 1, 1998 and Jan 1, 2000. Bottom: zoom-in to observations between 00:00:00 Jul 3, 1999 and 00:00:00 Jul 17, 1999.

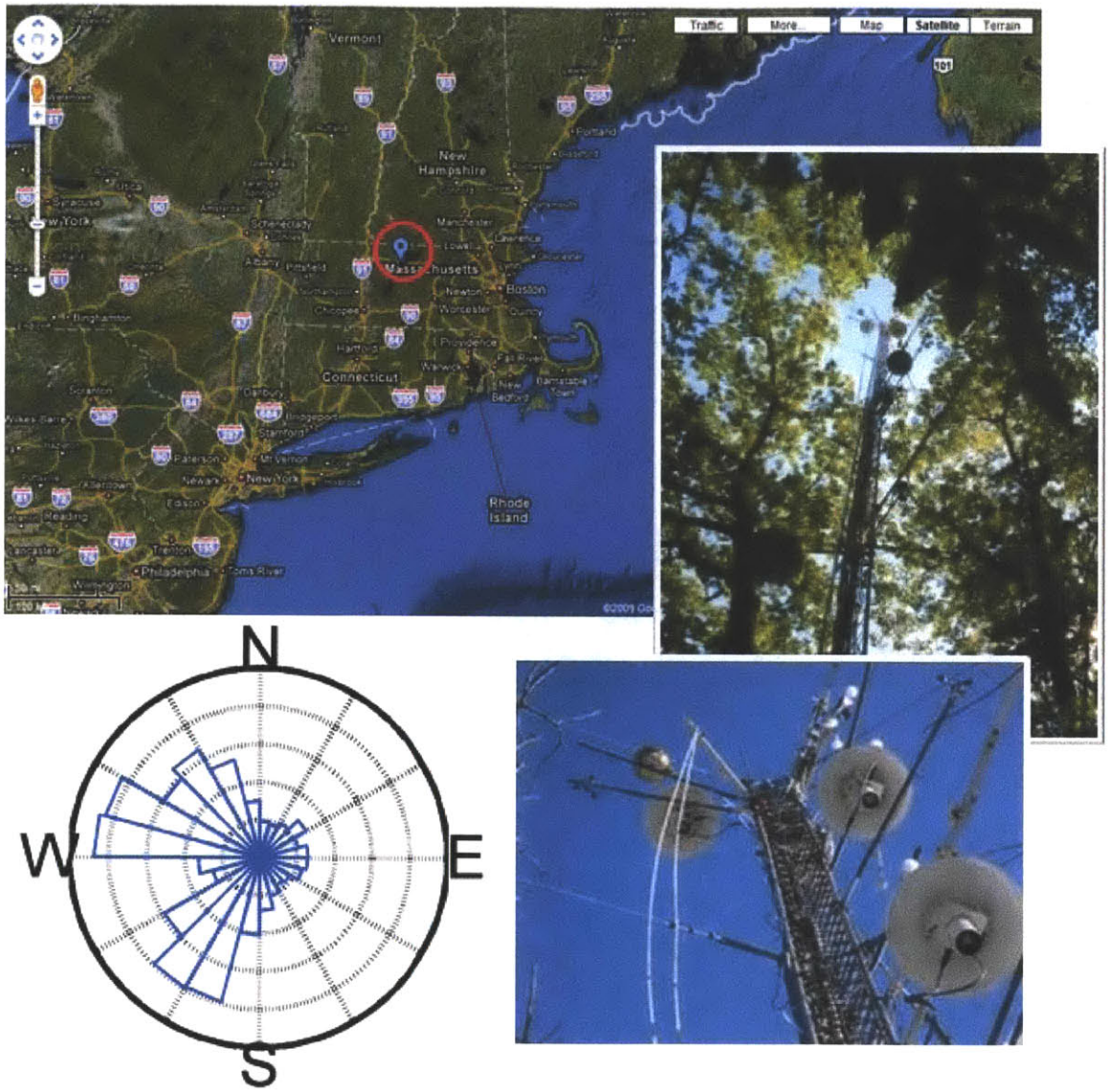

Figure 4-6: The Harvard Forest monitoring site. The map (courtesy of Google Maps) shows the location of the monitoring site in the Northeast industrial corridor. CO_2 measurements are made by sampling air from an eddy flux tower. The samples we used come from just above the canopy, ~30 meters above the ground. The eddy flux tower is shown in the photographs. Finally, a wind rose has been provided in the lower left to illustrate the dominant wind directions are from the west-northwest and soutwest.

Trends

Previous studies at Harvard Forest have made the following observations:

- CO_2 fluxes tend to vary interannually and seasonally. Interannual and decadal changes related to changes in forest composition, disturbances, and tree growth can be systematic but are difficult to capture through models [75]. Seasonality in net exchange is determined by the sum of respiration and photosynthesis. At Harvard Forest, photosynthesis is all but absent in the winter but dominates in the summer and is well explained by solar radiation [64]. Photosynthesis tends to dominate respiration on an annual time scale and fluxes are sensitive to weather and seasonality [76].

- Potosnak et al. [77] use a linear regression model[1] to estimate the contributions of fossil fuel combustion, regional variability, and local terrestrial processes to CO_2 concentrations. The estimated monthly means vary systematically by ~15 ppm, with lower values in the summer. According to their model, the bulk of diurnal variability in the summertime is due to regional biotic uptake and release rather than canopy scale terrestrial exchange. In Figure 3 of their paper, the authors present trends for several days in July 1995. The trends indicate a regional biotic influence of -20 to +5 ppm but a smaller local biotic influence ranging between -5 and 5 ppm. **This recommends the importance of accounting for regional variability as well as local biotic exchange.** Positive contributions from fossil fuel combustion tend to be about 4-5 ppm in the winter and 2-3 ppm in the summer. Barnes et al. [78] use measurements of CO and perchloroethylene (PCE) to demonstrate pollution events are strongly correlated with winds from the southwest, the direction of the D.C.-New York industrial corridor.

Figure 4-7 depicts CO_2 at Harvard Forest for three different time scales. The peak-to-peak amplitudes of the seasonal cycles do not conflict with the 15 ppm cycle estimated by Potosnak et al. [77], but not all of the variability in the data can be explained by a seasonal cycle of this magnitude. The short summertime period considered in the bottom plot in Figure 4-7 highlights the presence of significant diurnal cycles most likely due to photosynthesis and respiration, indicating that diurnal variability is an important factor in the overall variability of CO_2 at the site.

Figure 4-8 compares CO_2 data at Harvard Forest to contemporaneous data from Mauna Loa. Wintertime median CO_2 at Harvard Forest is consistently 10 ppm higher than at Mauna Loa. This systematic difference must be to the combined effects of regional respiration and fossil fuel emissions from the D.C.-New York corridor [77]. A much stronger seasonal cycle is apparent at Harvard Forest. The weaker cycle at

[1]The model they use is re-estimated on a monthly basis to capture seasonal variability. In each month, the regional mean is represented by a constant term. Deviations from the regional mean are modeled using (1) correlation between hourly mean CO and CO_2, to capture the effect of regional sources of CO_2 from fossil fuel combustion; (2) the ratio of hourly eddy covariance CO_2 flux to a term measuring the strength of vertical mixing; and (3) an empirical term representing the mean diurnal trend for CO_2 arriving at the canopy.

Mauna Loa is attributable to attenuation by atmospheric mixing. Both sites exhibit similar growing trends. Finally, the seasonal CO_2 trend at Mauna Loa is out of phase with that at Harvard Forest. We expect some lag due to the time required for atmospheric mixing to mix surface influences observed in the boundary layer into the free troposphere. From the data, it appears that the wintertime peak requires more time (~3-5 months) to mix than the summertime trough (~1-2 months), presumably the result of enhanced vertical mixing in the summer time.

Figure 4-9 plots time series of CO_2 and various meteorological variables for comparison. As expected, CO_2 concentrations tend to reach their minimum in the summer when sunlight is strongest. The CO_2 minimum appears to lag that of solar radiation very slightly, indicating the possible presence of a delay due to time required for biomass growth.

4.2.3 Salt Lake City

The Salt Lake CO_2 network is an urban atmospheric monitoring network located in Salt Lake City, Utah. The network is the longest running urban atmospheric monitoring network of which we are aware.

Salt Lake City is located in a valley surrounded by the Wasatch and Oquirrhs mountains to the east and west, respectively, and the Great Salt Lake to the northwest. Prevailing winds are from the north, although cross winds sometimes occur (personal communication with Steve Rowley).

The Salt Lake network consists of four urban sites and one rural site (see Figure 4-10). The first urban site (University) samples from the rooftop of a University of Utah building just east of downtown. The second (Residential) is located in a backyard in a residential neighborhood south of University by about 5 km. The third (Junior High) is located on the roof of a school approximately 10 km south of University. The fourth (Downtown) was initially located downtown on the roof of the X-Mission building but eventually moved westward to Rose Park; the data from these two locations has been merged into a single record. The rural site (Mine) is located to the southwest of the city on land owned by a copper mining company and is surrounded by agricultural land. The oldest site in the network is University, from which CO_2 data are available as far back as 2001.

Data

The data set we used is available to the public at `http://ecophys.utah.edu/download/CO2/`. Variables measured include hourly means and standard deviations of CO_2 concentration, wind velocity, air temperature, atmospheric pressure, relative humidity, and solar radiation. Total hourly rainfall is also recorded.

CO_2 concentrations at each site in the Salt Lake network are measured using infrared gas analyzers. The measurement method used has previously been described by Pataki et al. [79]. The infrared gas analyzers are high precision closed-path instruments which have been calibrated hourly using CO_2-free air and WMO standards to minimize errors due to span and zero drift. Two minute running averages are

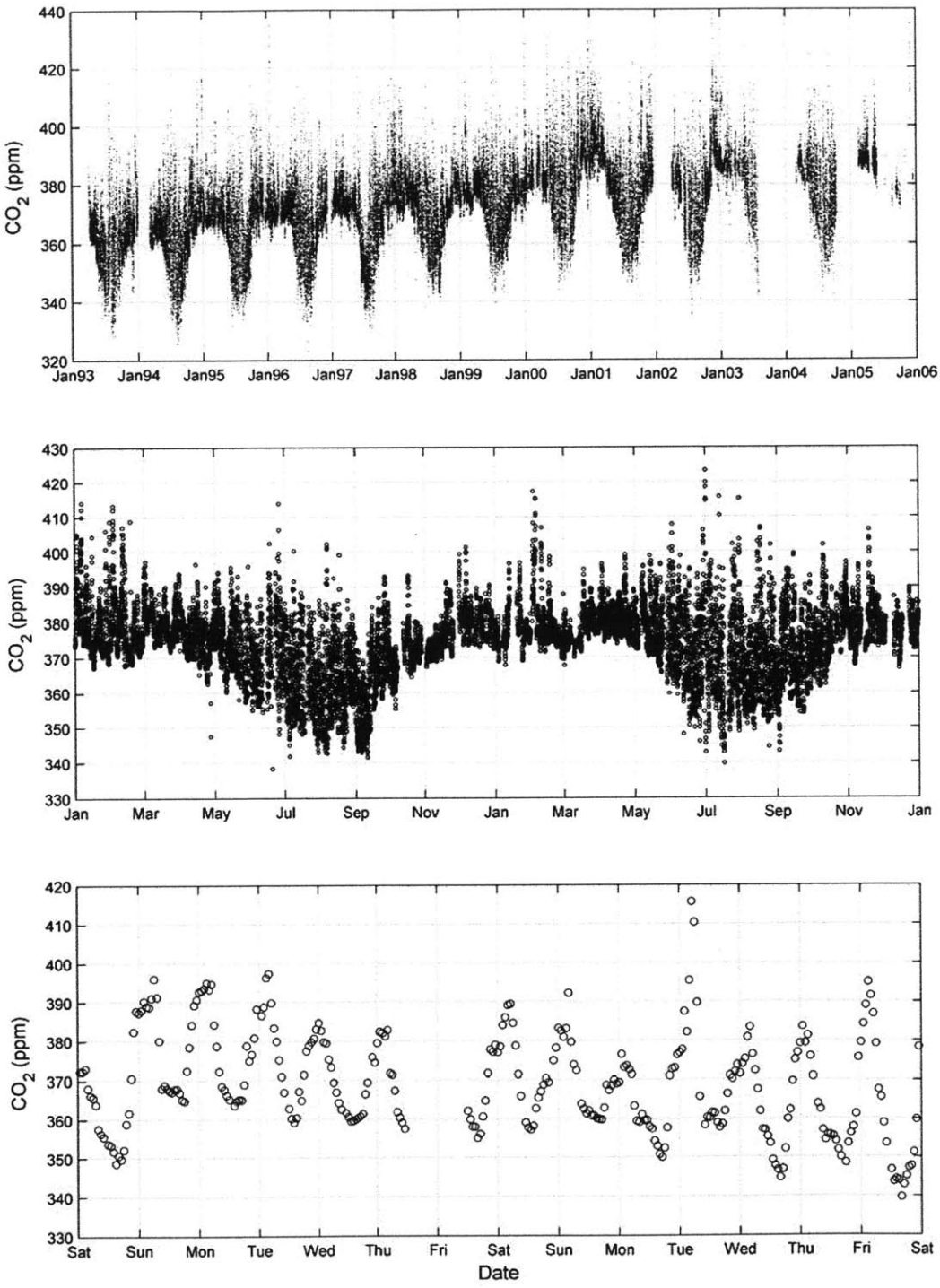

Figure 4-7: Atmospheric CO_2 trends at Harvard Forest's EMS. Top: complete hourly mean CO_2 time series from NIGEC. Middle: zoom-in to observations between Jan 1, 1998 and Jan 1, 2000. Bottom: zoom-in to observations between 00:00:00 Jul 3, 1999 and 00:00:00 Jul 17, 1999.

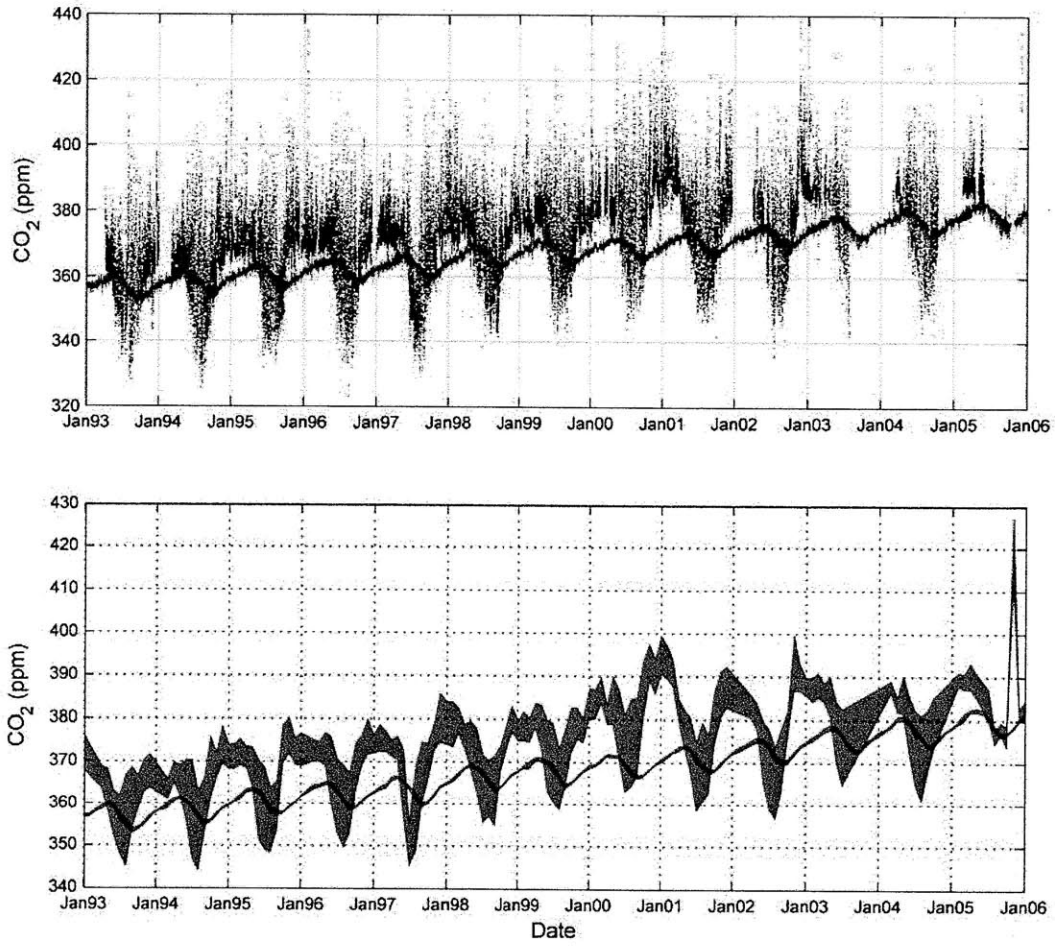

Figure 4-8: Atmospheric CO_2 at Harvard Forest (black) versus Mauna Loa (blue). Top: complete hourly mean CO_2 time series from NIGEC and NOAA. Bottom: the shaded regions represent the middle 50% of the hourly observations from Harvard Forest (grey) and Mauna Loa (blue). These regions were determined by computing the 25^{th} and 75^{th} percentiles of each for each month and shading the region between the resulting curves.

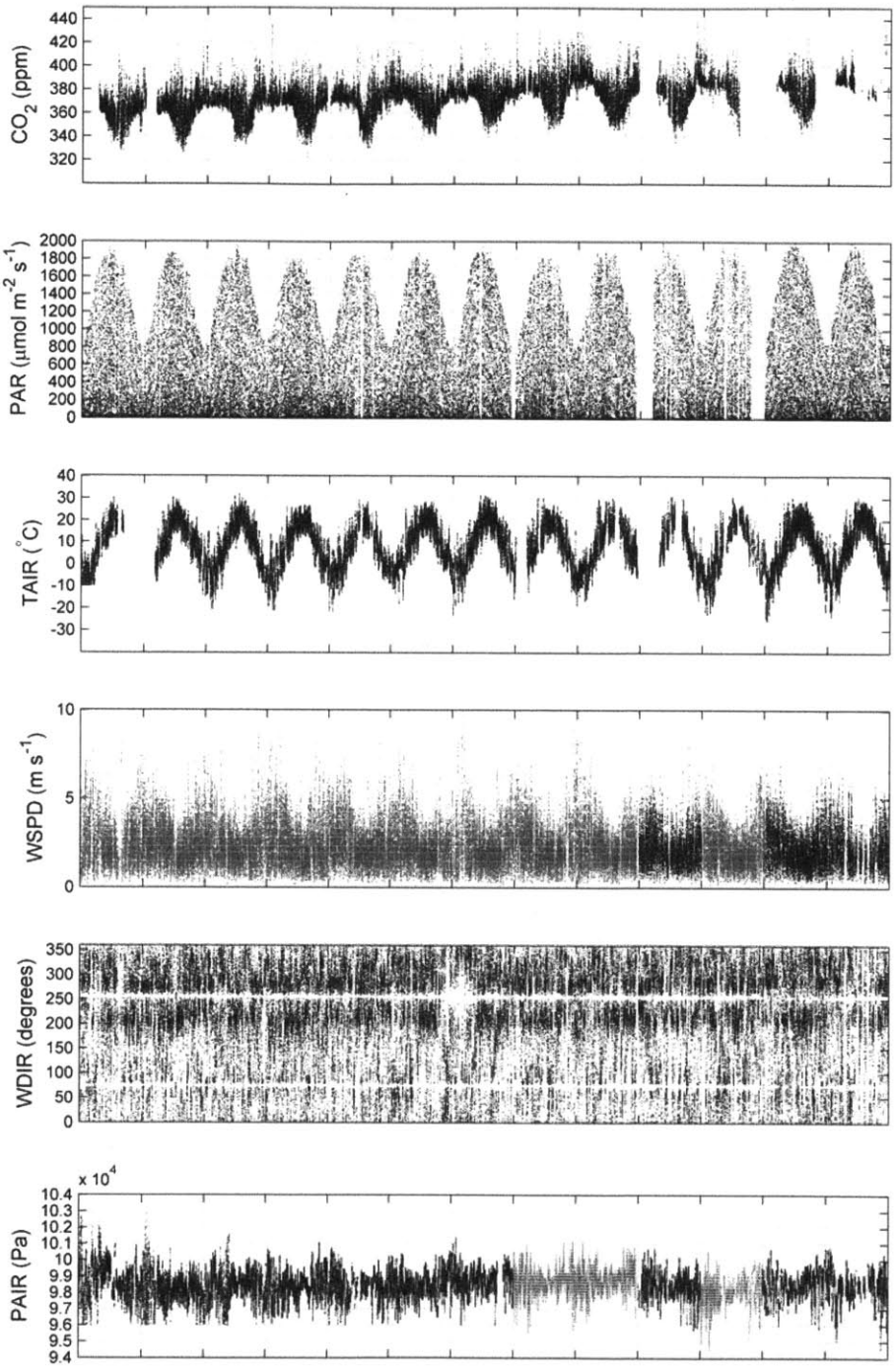

Figure 4-9: Relationships between atmospheric CO_2 at Harvard Forest and meteorological variables. For comparison to atmospheric CO_2 concentration, time series of the following are also plotted. PAR: photosynthetic active radiation; TAIR: air temperature measured by a precision thermistor; WSPD and WDIR: wind speed and direction, respectively, from a sonic anemometer; and PAIR: ambient air pressure. The time series for WSPD and PAIR appear striated in places; the reason for this is unknown. The white zones in WDIR are most likely the result of measurement deadzones.

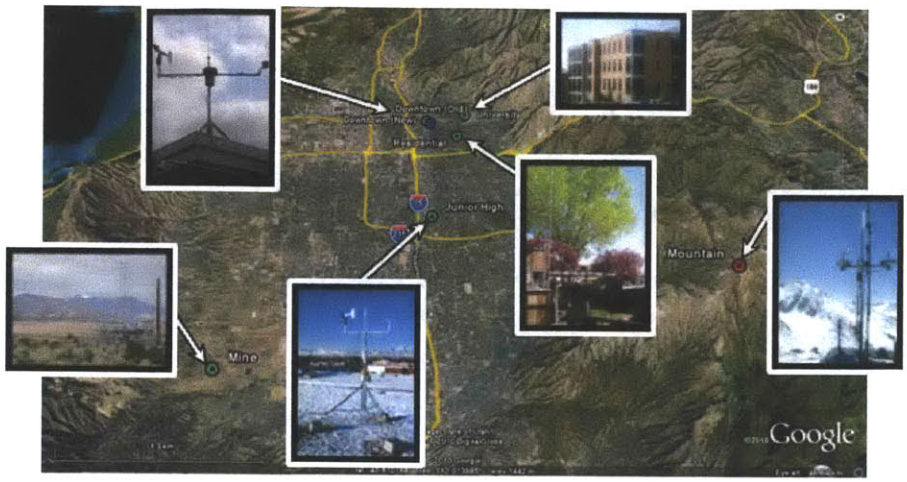

Figure 4-10: The Salt Lake network, map courtesy of Google Maps and photographs courtesy of the Ehleringer lab. Locations of active monitoring stations are shown in green; the blue marker indicates the old downtown site; the red marker is a background monitoring site that is part of the Rocky RACCOON monitoring network, which was not included in our analysis. The network is located between the Wasatch and Oquirrhs mountains to the east and west, respectively, and the Great Salt Lake to the northwest. The prevailing wind direction is from the north, although cross winds sometimes occur (personal communication with Steve Rowley).

recorded every five minutes. These five minute averages are then averaged to obtain hourly means and standard deviations. The hourly means are thus typically averages of 12 sub-hourly measurements.

Because most of our analysis focuses on the University site we often refer to it simply as "Utah". Unless stated otherwise, this refers to the University of Utah site.

Trends

Previous studies in Salt Lake have made the following observations:

- Measurements of $\delta^{13}C$ and $\delta^{18}O$ can be used to infer biotic and abiotic contributions to ambient CO_2 in urban settings [80, 81].

- CO_2 concentrations in Salt Lake City exhibit relatively small seasonal trends and larger diurnal fluctuations [82, 83, 81]. Diurnal enhancements can be >200-250 ppm, with greater variability in the winter when increased emissions coincide with reduced vertical mixing (see Figure 4-11).

- More than 60% of summertime diurnal variabilty is attributable to respiration, with a greater fraction at the Mine site than the urban locations. In the city, most wintertime variability is attributable to gasoline and natural gas combustion (see Figure 4-11).

- Nightime variability is far greater than daytime [82] (see Figure 4-12), the rationale being the same as that for the difference between winter and summertime variability.

- Contributions from natural gas vary significantly with time of day and are sensitive to changes in temperature [79]. Colder temperatures increase demand for natural gas which is used for heating.

- There are periods during which atmospheric stability leads to low lying inversions that trap emissions inside the valley, causing atmospheric pollutants to accumulate [83].

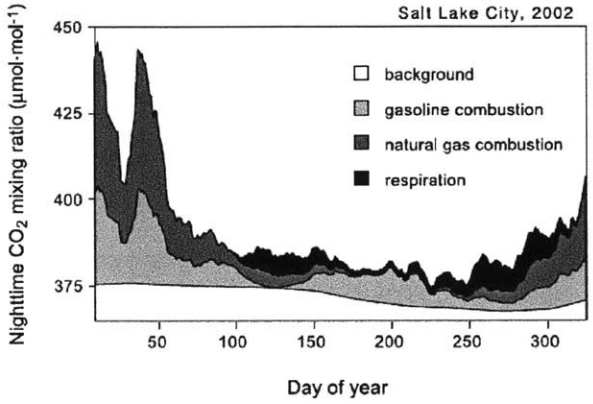

Figure 4-11: Contributions to CO_2 in Salt Lake City, adapted from Pataki et al. [82].

Figure 4-13 shows trends in CO_2 at University on three different time scales. A seasonal cycle is present, but its amplitude is difficult to assess by eye because of the substantial variability. CO_2 is lower in the summer as should be expected because of increased regional uptake by photosynthesis. The data support the notion that variability is greater in the winter than in the summer. Peak-to-peak diurnal fluctuations in the summer data considered range from 20-30 ppm. Short (1-2 week) upward shifts are apparent in the winter data as observed by Pataki et al. [82].

Figure 4-14 compares data from University and Mauna Loa. As at Harvard Forest, upward and seasonal trends are present in both and high frequency variability at University is far greater than that Mauna Loa. In the winter, CO_2 tends to be 10-15 ppm higher than at Mauna Loa, not dissimilar to what is observed at Harvard Forest. The data at Mauna Loa also appear to lag that at University in a similar manner as previously observed for Harvard.

Preliminary analysis of correlation between CO_2 from different sites in Salt Lake suggest higher correlations between sites on north-south than east-west transects, reflecting the dominant wind direction from the north. We find correlations between University and each of Residential and Junior High sites to be around 0.7. The correlation between hourly CO_2 at University and Downtown is approximately 0.5.

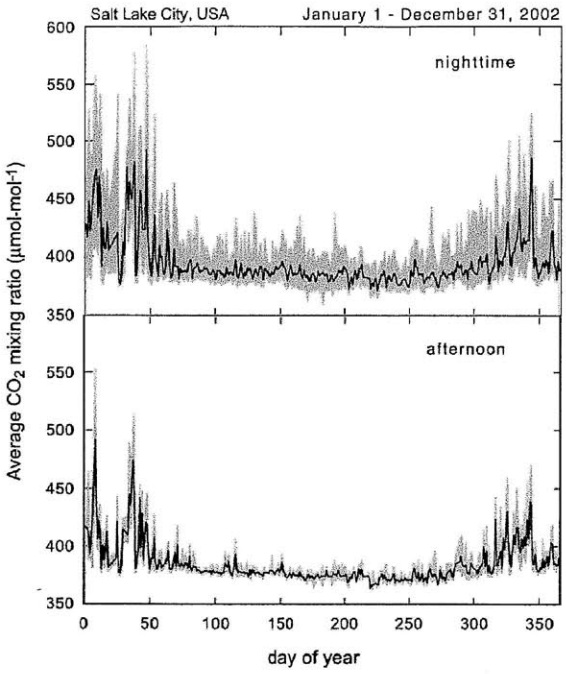

Figure 4-12: Daytime versus nighttime CO_2 variability in Salt Lake City, adapted from Pataki et al. [82].

Comparisons of CO_2 trends at University with those at Residential and Junior High support the notion that diurnal trends at these sites will at times be highly correlated.

4.3 Implications for modeling

Past studies of CO_2 variability and observations from the three sites considered here all confirm the presence of systematic interannual, seasonal and diurnal trends.

Most interannual variability is captured assuming a constant fraction of emissions from fossil fuel combustion remains airborne. The growing trend is currently ~2 ppm, and appears to be growing with increasing rates of fossil fuel consumption. Relatively small deviations occur due to processes that even relatively sophisticated models have difficulty capturing. Interannual trends are fairly consistent across sites, although absolute levels may differ because of the interhemispheric gradient.

Seasonal cycles are present, moreso in the Northern Hemisphere, which are due primarily to terrestrial photosynthesis and respiration. The magnitude of the seasonal cycle tends to grow larger north of the equator with increasing seasonal changes in sunlight and vegetation. The seasonal amplitude at Mauna Loa is 5 ppm. The amplitude in Barrow, Alaska is approximately 15 ppm. Seasonality is minimal south of the equator, 1 ppm at most. Systematic changes in seasonal amplitude have been observed at Mauna Loa due to changes in terrestrial sinks and wind patterns.

Diurnal deviations from seasonal trends can be larger in magnitude than the

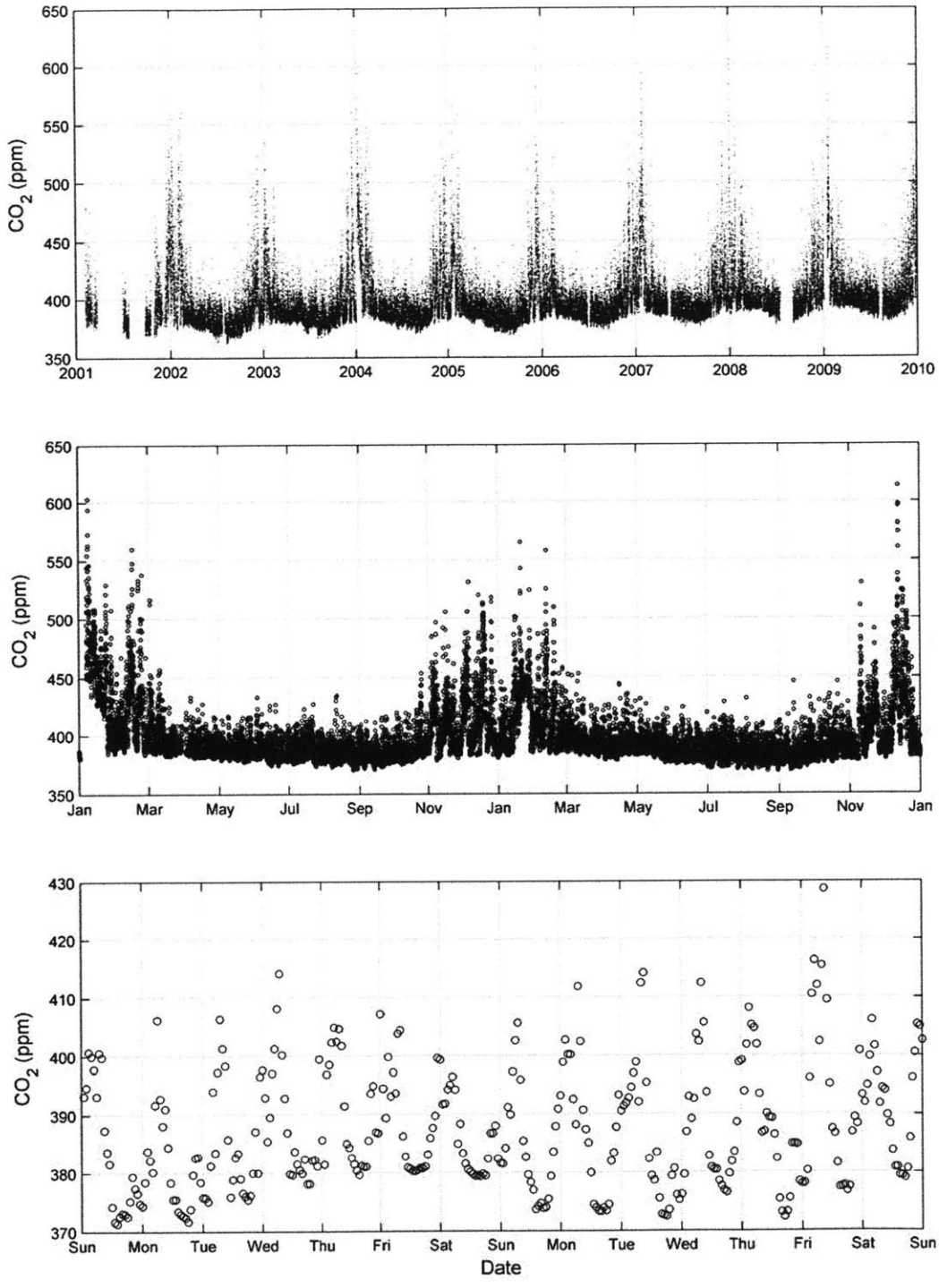

Figure 4-13: Atmospheric CO_2 trends at Salt Lake's University site. Top: complete hourly mean CO_2 time series from University of Utah. Middle: zoom-in to observations between Jan 1, 2004 and Jan 1, 2006. Bottom: zoom-in to observations between 00:00:00 Jul 3, 1999 and 00:00:00 Jul 17, 1999.

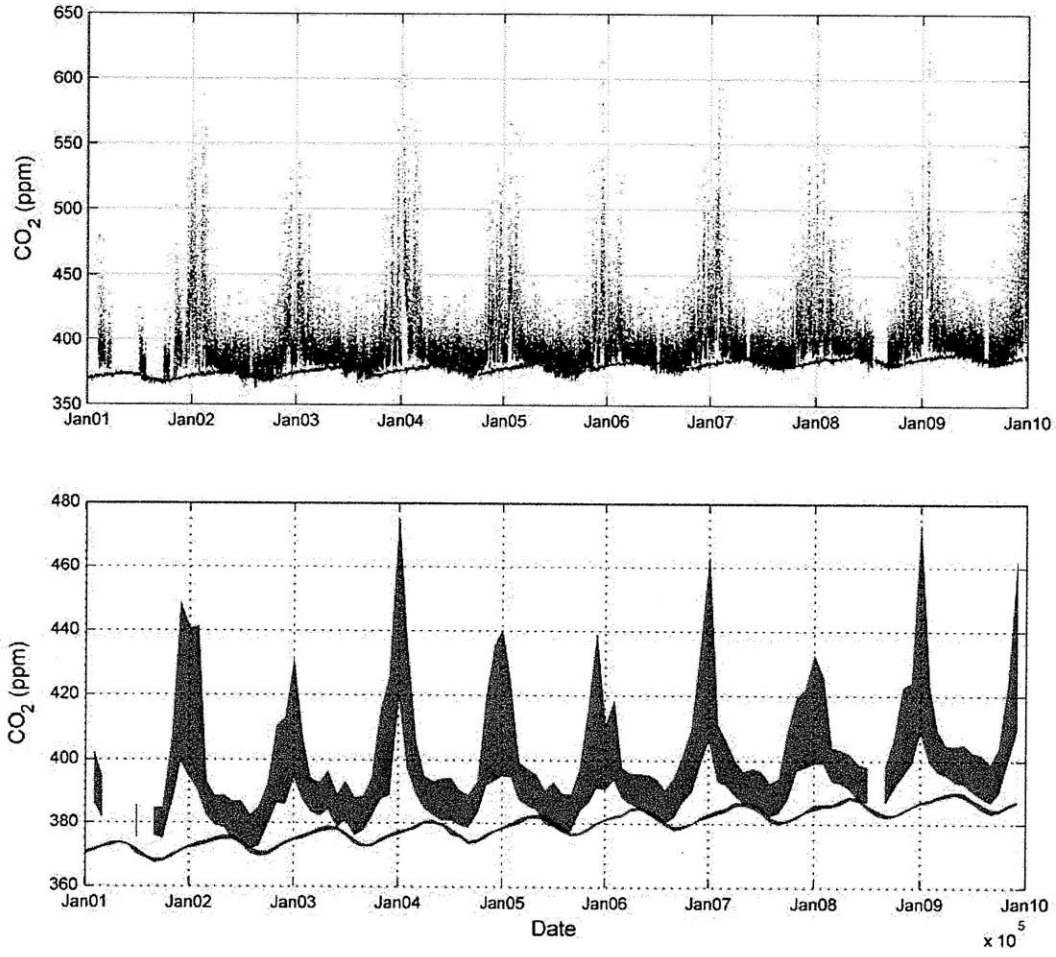

Figure 4-14: Atmospheric CO_2 at Salt Lake's University site (black) versus Mauna Loa (blue). Top: complete hourly mean CO_2 time series from University of Utah and NOAA. Bottom: the shaded regions represent the middle 50% of the hourly observations from University site (grey) and Mauna Loa (blue). These regions were determined by computing the 25^{th} and 75^{th} percentiles of each for each month and shading the region between the resulting curves.

seasonal cycles themselves. Both local, canopy (~1 km) scale processes as well as advection of regional (1 to 100+ km) air masses affect diurnal variability. Diurnal trends therefore depend not only on the direct vicinity of the sensor but also upon the regional context. At Mauna Loa, local influences are minimal except occasionally at night when downslope winds bring CO_2 emitted from the volcano's summit. At Harvard Forest, both regional and local influences are important with regional biotic contributions tending to dominate. Combustion emissions from hundreds of kilometers upwind may contribute 5 ppm enhancements at Harvard Forest in the winter. In Salt Lake City, diurnal variability is dominated by fossil fuel combustion, but respiration does contribute significantly in the summer. The scale of diurnal variability in Salt Lake is much greater than at Harvard Forest, presumably due to larger scale emissions.

Chapter 5

Detrending models for atmospheric CO_2

The previous chapter describes the types of trends observed in atmospheric CO_2 data and where the trends come from. This chapter uses this knowledge to build several physical models for detrending atmospheric CO_2 data. A key question is, which model should be used for detrending? The processes affecting atmospheric CO_2 are numerous and complex, so it is difficult to objectively know when to stop adding terms to the model. Deviance information criterion is introduced as a Bayesian method for choosing between models that avoids overfitting. The method is demonstrated for the set of models developed in the beginning of the chapter. The final model selected will be used later in Chapter 7 to detect urban shifts in atmospheric CO_2.

5.1 Developing models

5.1.1 Global emissions

Past data indicate the presence of an increasing trend related to emissions from fossil fuel combustion. The use of emissions and surrogates for emissions (e.g., exponential functions [61]) is commonplace in models of atmospheric CO_2. Figure 5-1 illustrates trends in net emissions and Mauna Loa atmospheric CO_2 data side by side for comparison. The shapes of the trends are very similar, supporting the hypothesis that emissions drive global increases in atmospheric CO_2 over time. If a constant fraction of emissions remains airborne, then interannual increases in CO_2 can be explained by a model of the form

$$CO_2^{inter}(t) = CO_2(t_0) + f \cdot [E(t) - E(t_0)] \tag{5.1}$$

where t_0 is a reference time and t the time of interest, CO_2 is the atmospheric carbon dioxide concentration, $E(t) - E(t_0)$ is the net emissions from fossil fuel combustion since time t_0, and f is the fraction of the emissions remaining airborne. The unknowns are the model parameters $CO_2(t_0)$ and f. Since the model is linear, ordinary least squares can be used to estimate the parameters from concentration and emissions

data. For example, using the annual Mauna Loa data shown in Figure 5-1, least squares estimates the global mean concentration of CO_2 in 1959 to be approximately 316.5±0.2 ppm, with an emissions contribution of $7.5\times10^{-2}\pm0.04\times10^{-2}$ ppm/$GtCO_2$. The least squares fit is shown in Figure 5-2. From the figure, it is apparent that the decadal trend in global mean atmospheric CO_2 correlates well with net emissions from fossil fuel combustion.

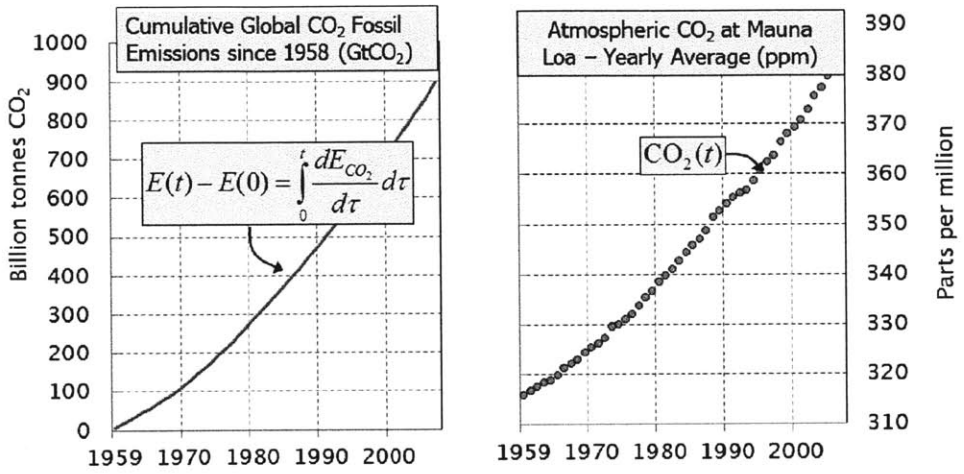

Figure 5-1: Global emissions and atmospheric CO_2. The data on the left come from CDIAC and represent annual global emissions of CO_2 from fossil fuel combustion. The data on the right are yearly atmospheric CO_2 data from Mauna Loa.

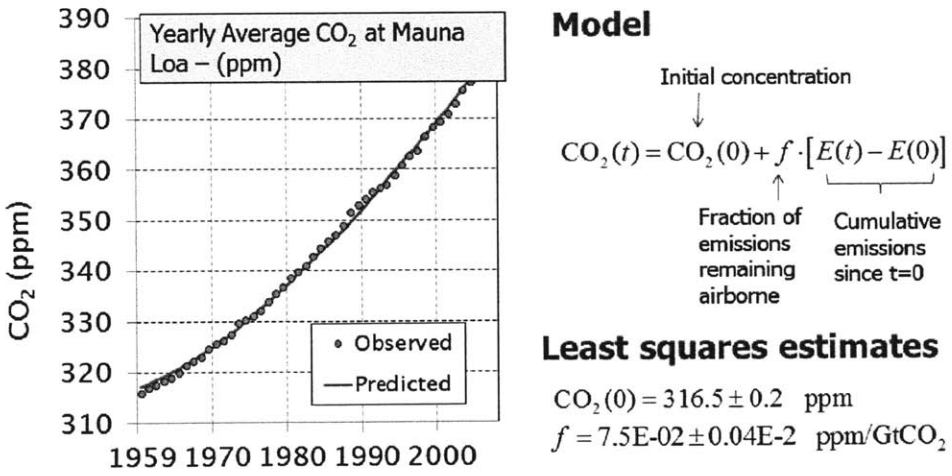

Figure 5-2: Fitting global emissions to yearly atmospheric CO_2 data at Mauna Loa.

Next, CO_2 data from Harvard Forest were fitted using emissions data from the U.S. Energy Information Administration[1].

[1]The choice of the new emissions data source reflected the order in which analysis was performed.

The EIA reports annual global emissions since 1980. For modeling, the data for each year has been dated to the end of the reported year (i.e., 12:59:00 Dec 31). The result is a series of annual emissions $\Delta E(t_i)$, where $i = 1980, 1981, ...$ Net emissions $E(t_i)$ are obtained by summing the annual emissions, as in $E(t_n) = \sum_{i=1}^{n} \Delta E(t_i)$. Table 5.1 presents the annual and net annual emissions data used here. Assuming fossil fuel emissions of CO_2 to be relatively constant throughout the year, the net emissions at some time t has been obtained by linearly interpolating the curve defined by the $E(t_i)$'s. The result is a curve that accelerates upward in time, as shown in Figure 5-3. This curve is the basis for the emissions models in the rest of this chapter.

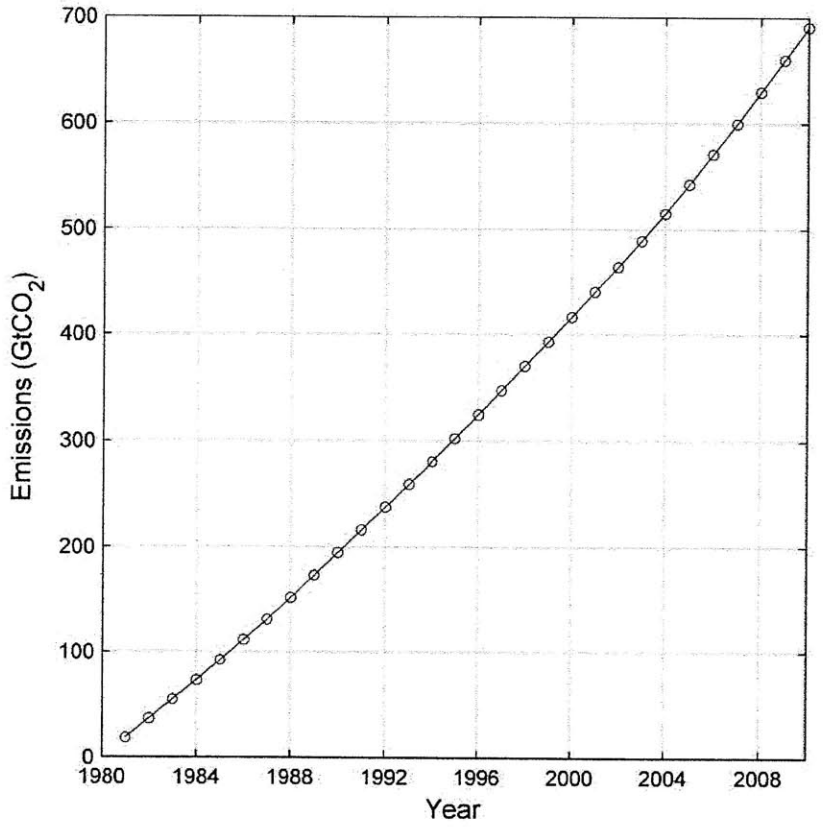

Figure 5-3: Cumulative international fossil fuel emissions since 1980 based on data from the U.S. Energy Information Administration.

Figure 5-4 depicts the emissions model fit alongside monthly averages of the original (hourly) Harvard Forest data. The emissions model is clearly incapable of capturing seasonal fluctuations in the data. Much of the seasonal variation should be due to

Although slight differences do exist between the datasets, this was not a motivating factor in our choice of emissions data. In future studies, the model choice methods presented later could be used to objectively choose between two or more sources of emissions data.

Table 5.1: Fossil fuel emissions data

Date	Annual emissions ($GtCO_2$ yr^{-1})	Net emissions ($GtCO_2$)
31-Dec-80	18.4	18.4
31-Dec-81	18.2	36.6
31-Dec-82	18.1	54.7
31-Dec-83	18.2	73.0
31-Dec-84	19.1	92.1
31-Dec-85	19.5	111.6
31-Dec-86	19.9	131.6
31-Dec-87	20.5	152.0
31-Dec-88	21.2	173.2
31-Dec-89	21.5	194.7
31-Dec-90	21.6	216.3
31-Dec-91	21.5	237.8
31-Dec-92	21.4	259.2
31-Dec-93	21.6	280.9
31-Dec-94	21.8	302.7
31-Dec-95	22.2	324.8
31-Dec-96	22.7	347.5
31-Dec-97	23.1	370.5
31-Dec-98	23.0	393.5
31-Dec-99	23.3	416.9
31-Dec-00	23.8	440.7
31-Dec-01	23.9	464.6
31-Dec-02	24.7	489.3
31-Dec-03	25.9	515.2
31-Dec-04	27.5	542.7
31-Dec-05	28.4	571.1
31-Dec-06	28.9	600.0
31-Dec-07	29.8	629.8
31-Dec-08	30.5	660.3
31-Dec-09	30.4	690.7

photosynthesis and respiration. These biological processes depend on solar radiation. The next model considered therefore uses solar radiation to explain seasonal changes in atmospheric CO_2.

Emissions Model $\mathcal{M}_E$

$$CO_2(t) = CO_2(0) + f \cdot [E(t) - E(0)]$$

Least squares estimates

$CO_2(0) = 362.8 \pm 0.2$ ppm

$f = 8.6\text{E-}02 \pm 0.1\text{E-}2$ ppm/GtCO_2

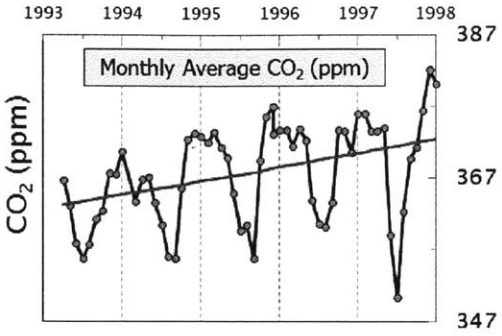

Figure 5-4: Emissions fit to atmospheric CO_2 data at Harvard Forest.

5.1.2 Solar radiation

The amount of solar radiation reaching the Earth changes seasonally for locations away from the equator because the Earth is tilted relative to its plane of orbit around the Sun. The tilt, which is called the declination angle ($\angle\delta$), varies between $\pm 23.44°$ and can be modeled as

$$\angle\delta(t) = 23.44° \cdot \cos\left[2\pi \frac{\text{DayOfYear}(t) + 10}{365}\right] \tag{5.2}$$

where DayOfYear(t) is the number of days (0-365) elapsed since 00:00:00 Jan 1 of the year containing time t.

At solar noon, the flux of radiation normal to a point on the Earth's surface depends on the difference between the declination angle and the angle between the local normal vector and the equatorial plane, with the latter angle being simply the latitude of the monitoring site. The difference between these angles is called the zenith angle. The zenith angle is location dependent, and as already described is determined as

$$\theta_z^{\min}(latitude, t) = latitude - \angle\delta(t) \tag{5.3}$$

The noontime solar flux $SOLRAD(latitude, t)$ is determined here as

$$SOLRAD(latitude, t) = I_{sol} \cdot \max\left\{0, \cos\left[\theta_z^{\min}(latitude, t)\right]\right\}, \tag{5.4}$$

where $I_{sol} \approx 1,367$ Wm^{-2} is the typical flux of solar radiation reaching the Earth's outer atmosphere[2] throughout the year, and $\theta_z^{\min}$ is the local zenith angle.

The hypothesis is that atmospheric CO_2 is negatively correlated with solar radi-

[2]Commonly referred to as the "solar constant".

ation, so the model is

$$CO_2^{seas}(t) = -Y_R \cdot SOLRAD(latitude, t) \tag{5.5}$$

where Y_R is the yield (ppm/W m^{-2}) of atmospheric CO_2 per flux of solar radiation.

Figure 5-6 plots solar radiation flux reaching the outer atmosphere and atmospheric CO_2 for comparison. There is a clear lag between the decrease in atmospheric CO_2 and the increase in solar radiation. This lag is likely due to time required for vegetation to grow. To account for this lag, a lag time can be introduced into the solar radiation function. The resulting model is

$$CO_2^{seas}(t) = -Y_R \cdot SOLRAD(latitude, t - \tau_{growth}) \tag{5.6}$$

Since the model is nonlinear, nonlinear least squares is used to estimate the parameters[3]. Figure 5-5 plots the fit to the Harvard data obtained using solar radiation. The seasonal changes are poorly captured by the model. This is due to asymmetry in the seasonal trend in atmospheric CO_2, as highlighted in Figure 5-6. The problem is, the lag shown in the latter figure is not the same in the autumn as in the spring. It is impossible to accurately capture the seasonal fluctuation in atmospheric CO_2 by simply shifting the solar radiation curve to the right.

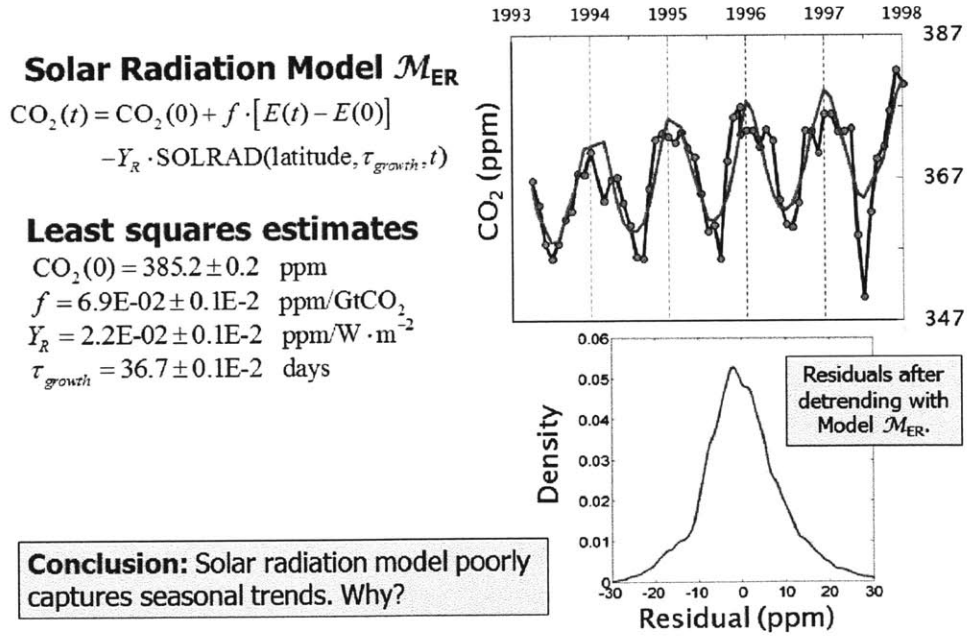

Figure 5-5: Solar fit to atmospheric CO_2 data at Harvard Forest. The model includes both emissions and solar terms.

[3]A more careful approach should employ Bayesian estimation; since the goal of this work is primarily demonstration of ideas, least squares has been used instead as it entails less computational effort.

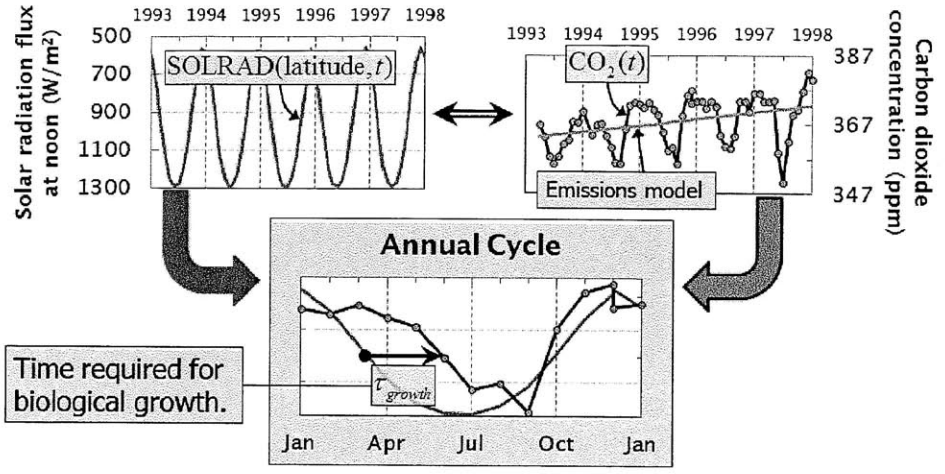

Figure 5-6: Solar radiation and atmospheric CO_2. Seasonal changes in atmospheric CO_2 and solar radiation are compared. CO_2 lags solar radiation in the spring by a time τ_{growth} related to time required for biological growth.

5.1.3 Leaf area

An alternative to using solar radiation is to use knowledge of seasonal changes in biomass to predict CO_2. When atmospheric CO_2 concentration drops in the summer, it does so because carbon dioxide molecules are being utilized for plant growth. One place where the CO_2 goes is into leaves. A logical next step in the modeling is to see how well changes in leaf mass can predict seasonal fluctuations in atmospheric CO_2.

Some sites, like Harvard Forest, measure a quantity related to leaf mass called the leaf area index (LAI). Leaf area index is leaf area per unit soil area. Obviously the greater the LAI, the greater the amount of CO_2 that has been removed from the atmosphere. A model relating seasonal changes in CO_2 to LAI can thus be conceived as

$$CO_2^{seas}(t) = -Y_L \cdot LAI(t, \boldsymbol{\alpha}_{DLM}) \tag{5.7}$$

where Y_L is the yield (in ppm) of atmospheric CO_2 per unit leaf area index and $\boldsymbol{\alpha}_{DLM}$ is a vector of five DLM parameters which will be defined shortly.

A dynamic leaf model (DLM) proposed by Dickinson et al. [84] and Liu et al. [42] has been used to simulate leaf area. The authors used their dynamic leaf model as a way to obtain smooth leaf area estimates when inverting satellite albedo measurements. Soil moisture and air temperature govern leaf dynamics in the general version of the model. As Liu et al. point out, air temperature is more important in cold weather, deciduous environments like Harvard Forest. Since Harvard Forest and Salt Lake are both situated in colder climates, a simpler version which accounting only for air temperature was chosen for use.

In the DLM, air temperature (T, in Kelvin) affects leaf growth and abscission in

a switch-like manner through a Heaviside function[4]

$$R(x) = [1 + \exp(-2x)]^{-1} \tag{5.8}$$

where $x = x(T)$ is a normalized temperature given by

$$x = \frac{T - T_{\min}}{\Delta T} \tag{5.9}$$

When temperature is low, R is close to zero and leaf decay dominates growth. The reverse is true at high temperatures, with R providing switch like behavior to the leaf dynamics. Dickinson et al. [84] note that such leaf area models tend to be prone to greater uncertainty in the spring and autumn, when leaf area changes most. This reflects uncertainty in the timing of leaf growth and abscission.

Leaf area index is modeled by

$$\frac{dLAI}{dt} = -\lambda(LAI) \cdot LAI \tag{5.10}$$

where λ is the inverse time scale (in days) of leaf growth (negative) or decay (positive). λ is given by

$$-\lambda(LAI) = \underbrace{\lambda_0 \{1 + a[1 - R(x)]\}}_{\text{Specific leaf loss rate}} - \underbrace{\lambda_0 \left\{ R(x) \left(\frac{LAI_{\max}}{LAI} \right) \left[1 - e^{-LAI} \right] \right\}}_{\text{Specific leaf growth rate}} \tag{5.11}$$

The unknown model parameters are the scaling coefficient Y_L and the DLM parameters λ_0, a, LAI_{max}, T_{min} (in Kelvin), and ΔT (in Kelvin). Since the model is nonlinear, nonlinear least squares is used to estimate all six parameters. Due to the presence of the differential equation, forward sensitivity analysis is used in the optimization procedure.

Figure 5-7 shows the dynamic leaf model (DLM) fit to monthly average CO_2 data from Harvard Forest. The model captures the seasonal fluctuations better than the solar radiation model did. Figure 5-8 shows the power spectrum of the residuals from the leaf model. The power spectrum contains no statistically significant peaks below 10 cycles per year, indicating that all long term and seasonal signals have been removed by the model.

5.1.4 An extended solar radiation model

While leaf area index does a good job capturing seasonal changes in the data, few sites measure LAI. The solar radiation model might be considered as a substitute for the more complex leaf model. The issue with the solar radiation model is that because it is incapable of capturing the changing lag between solar radiation and atmospheric CO_2 during the year the residuals that result contain a semiannual signal, as highlighted in

[4] $R \to 0$ for large negative x and $\to 1$ for the opposite.

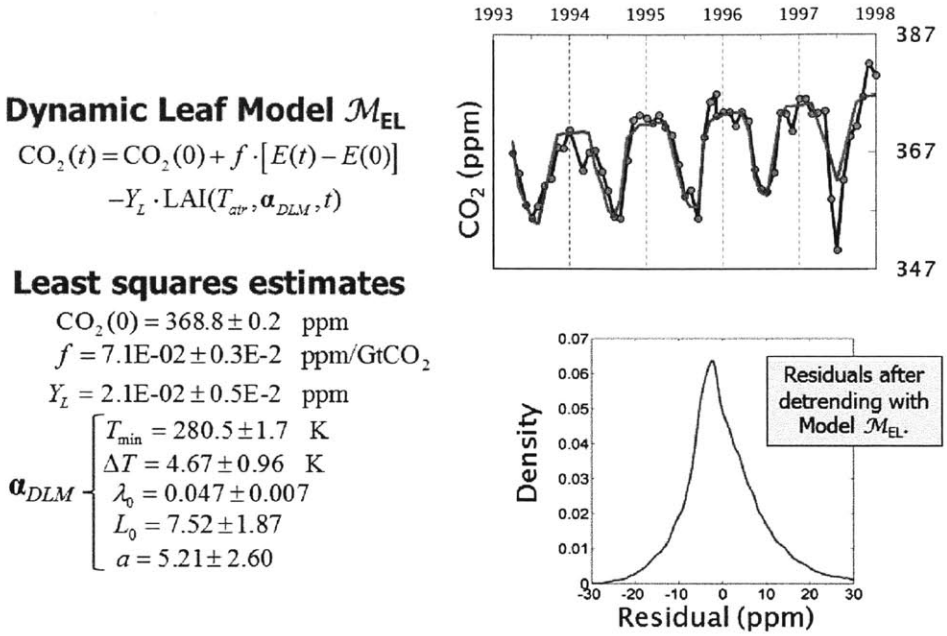

Figure 5-7: Leaf model fit to atmospheric CO_2 data at Harvard Forest. The model includes both emissions and leaf terms.

the upper plot of Figure 5-9. To remove this semiannual signal, a 6-month harmonic term can be added to the model. The result is an extended solar radiation model of the form

$$\begin{aligned} \mathrm{CO}_2^{seas}(t) = & - Y_R \cdot SOLRAD(latitude, t - \tau_{growth}) \\ & + a_1 \sin\left(4\pi \cdot \frac{\mathrm{DayOfYear}(t)}{365}\right) + a_2 \cos\left(4\pi \cdot \frac{\mathrm{DayOfYear}(t)}{365}\right) \end{aligned} \tag{5.12}$$

where DayOfYear(t) is the number of decimal days elapsed since the beginning of the year to which time t belongs. Adding the 6-month harmonic to the solar radiation model eliminates the semiannual signal remaining when the solar radiation model is used to remove seasonality from the Harvard data. This is demonstrated by the lower plot in Figure 5-9.

The origin of the harmonic term introduced here is complex and site dependent. Previous studies have used it to explain deviations from emissions predicted trends. Such deviations have been noted to occur because of changes in transport and upwind terrestrial exchange [63]. Other reasons have also been hypothesized [61]. Many studies have used a 6-month sinusoidal term identical to the one used here to model trends in atmospheric CO_2 [69, 85, 86, 87].

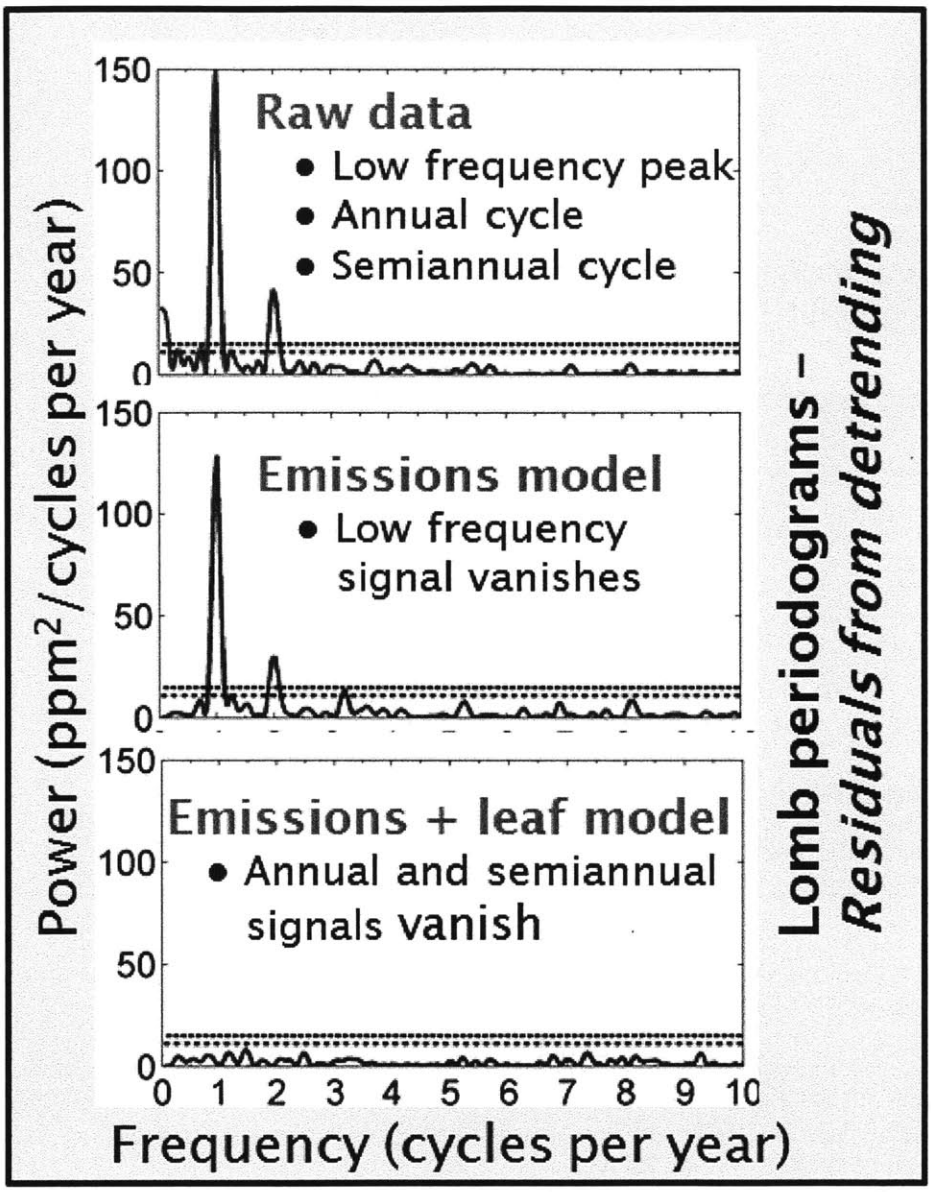

Figure 5-8: Three Lomb periodograms are shown for, from top to bottom, (1) the raw data, (2) the residuals obtained by removing emissions driven growth, and (3) the residuals after removing both emissions and leaf trends. The horizontal dashed lines are thresholds corresponding to 0.1% (top) and 5% levels of significance.

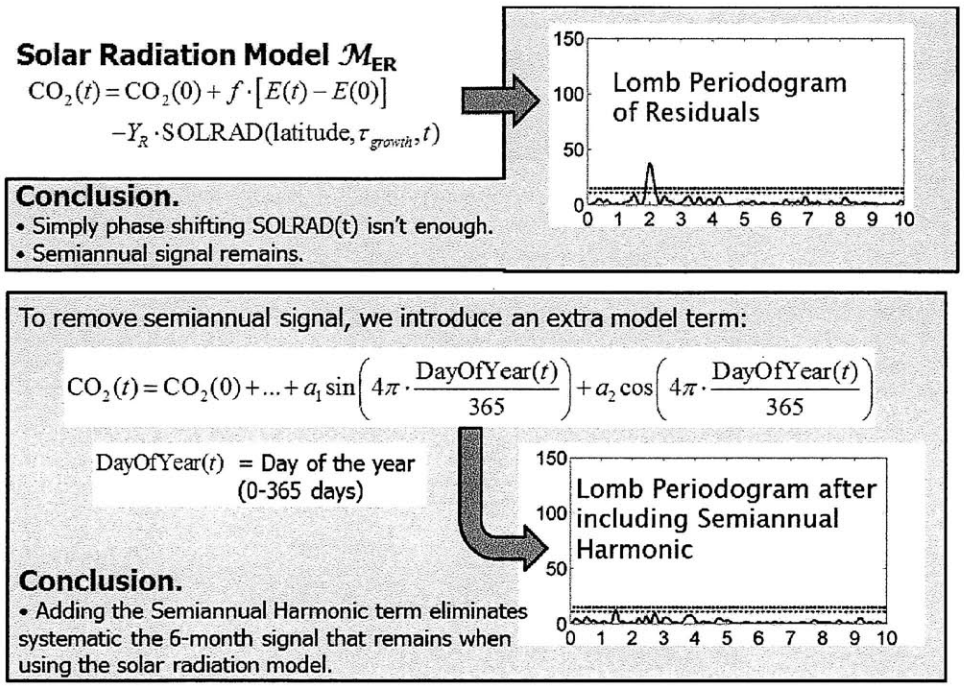

Figure 5-9: Extending the solar radiation model. Not all sites are equipped to measure leaf area. The idea illustrated in this figure is that the seasonal signal in the original data can be eliminated by extending the solar radiation model using sine functions. The resulting model is much simpler to simulate than the leaf model and does not require leaf data for validation. The plots shown are Lomb periodograms of the residuals from the solar radiation model (top) and for the extended version of the same model (bottom). The horizontal dashed lines in each plot are thresholds corresponding to 0.1% (top) and 5% levels of significance.

5.1.5 Slope-intercept

Over short periods of time, the emissions trend can be approximated by a simple linear trend. Thus, models of the form

$$CO_2^{inter}(t) = CO_2(t_0) + CO_2' \cdot (t - t_0), \tag{5.13}$$

where CO_2' is a new model parameter (with units of ppm/day) representing the average rate of change in atmospheric CO_2, might also been considered. The suitability of such a model will depend upon the size of the training window. This type of model has been used before to model long term trends in CO_2 [69, 85, 87].

5.1.6 Summary of proposed models

Different combinations of the terms just described can be used to formulate different models for atmospheric CO_2. Any model that is proposed should at minimum include a term explaining the growing trend in atmospheric CO_2. Over short periods of time, slope-intercept terms can be used for this purpose; longer periods of time may necessitate the use of emissions data for describing nonlinear trends in the data. A seasonal term composed of either leaf model predictions or solar radiation can be added to either of these growing terms, with the optional addition of a harmonic term to account for missing low frequency terms. The final models considered here thus represent the sum of interannual and seasonal components as in

$$CO_2(t) = CO_2^{inter}(t) + CO_2^{seas}(t) \tag{5.14}$$

The interannual terms include the slope-intercept (S) model and the emissions (E) model. Seasonal terms include solar radiation (S), dynamic leaf area index (L), and the semiannual harmonic term (O) used earlier to extend the solar model. Table 5.2 summarizes the model terms developed earlier.

Table 5.2: Model terms

Model	Expression	Notes
S	$\theta_1 + \theta_2(t - t_0)$	Slope-intercept term. Explains interannual trend as straight line. This type of model has been used before to explain long term trends in CO_2 [69, 85, 87].
E	$\theta_1 + \theta_2[E(t) - E(t_0)]$	Cumulative international emissions since Jan 1980. $E(t)$ has been formed by interpolating $E(t_i)$. Annual emissions data was obtained from the U.S. Energy Information Administration, so the t_i's were dated to 12:59:00 Dec 31 of the reported year and the cumulative emissions value at the end of the n^{th} year was obtained as $E(t_n) = \sum_{i=1}^{n} \Delta E(t_i)$. The use of emissions and surrogates for emissions (e.g., exponential functions [61]) is commonplace in models of atmospheric CO_2.
R	$-\theta_3 SOLRAD(t - \theta_4)$	Lagged solar radiation term. By restricting θ_3 to be positive, the model has been forced to correlate high values of solar radiation with low values of CO_2. The lag parameter θ_4 represents the time required for the acceleration of biological growth.
L	$-\theta_3 LAI(t, \theta_4..\theta_8)$	Dynamic leaf term. Improves on the solar radiation model by correlating CO_2 to leaf mass as predicted by the dynamic leaf model described by Dickinson et al. [84] and Liu et al. [42]. The form of the model used here predicts leaf mass changes with air temperature. This version of the model applies to cold weather deciduous forests [42]. Temperature increases have been cited as a possible reason for observed increases in amplitudes of seasonal trends in CO_2 [10]. Parameters $(\theta_4...\theta_8) = (T_{min}, \Delta T, \lambda_0, L_0, a)$
O	$\theta_9 \sin[4\pi(t - t_0)/365] + ...$ $\theta_{10} \sin[4\pi(t - t_0)/365]$	The origin of this term is complex and may be site dependent. Here it is used to remove systematic signals remaining because of poor fit. Past studies have linked this type of term to slow systematic deviations from emissions predicted trends. Such deviations have been noted to occur because of changes in transport and upwind terrestrial exchange [63]. Other reasons have also been hypothesized [61]. Many previous studies have used this term to model trends in atmospheric CO_2 [69, 85, 86, 87].

Models formed by combining individual interannual and seasonal terms are named by the letters designating the individual components. For example, the model combining slope-intercept trend with a dynamic leaf term would be designated "SL". In building models, each model has been required to have one interannual term, either S or E. The harmonic term (O) is used to diagnose lack of fit, so the solar radiation and leaf terms (S and L) have been given higher priority than. A total of ten possible models are thus entertained: S, E, SR, ER, SL, EL, SRO, ERO, SLO, and ELO. At Mauna Loa leaf area models are not considered because temperature data are unavailable in the CO_2 data provided by NOAA, so six models were possible at that site.

5.1.7 Improving the model

Figure 5-10 illustrates the effect improving the model has on the distribution of the residuals at Harvard Forest. As the model is improved, the distribution narrows slightly. This reflects the model's ability to capture systematic growing and seasonal trends in the data. Significant variability remains in the residuals because of diurnal variability.

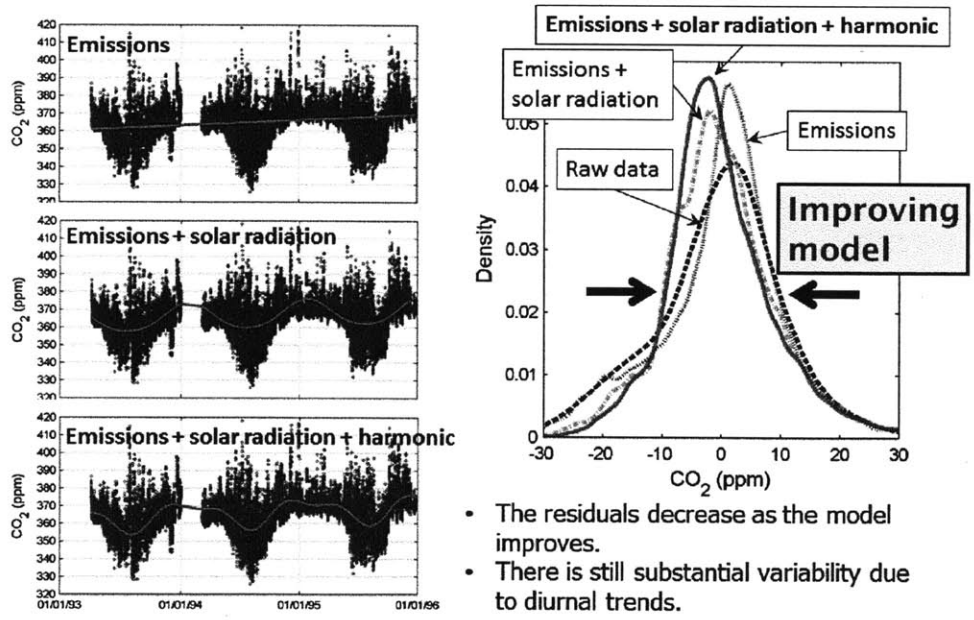

Figure 5-10: The effect of improving the detrending model is shown for hourly CO_2 data from Harvard Forest. The figures on the left are least squares fits obtained using emissions, solar radiation, and extended solar radiation models. The distributions of the residuals are compared on the right.

While for shift detection decreasing the variance of the residuals is desirable because it increases the signal to noise ratio, there is nothing to stop an operator from adding terms to the model until it fits the data perfectly. The next section introduces the deviance information criterion (DIC) as a way to avoid overfitting.

5.2 Choosing a model

We used three measures to examine model performance and choose the best model for each observation site. The metrics used included:

- Sum of squared residuals (RSS).
- Akaike's information criterion (AIC).
- Deviance information criterion (DIC).

Choosing the model that minimizes the sum of squared residuals leads to overfitting. AIC avoids overfitting by penalizing excessive numbers of parameters in the models proposed, but is based on the frequentist assumption that parameters are exact. DIC is a Bayesian alternative that is much easier to compute than Bayes factors but accounts for uncertainty in the parameters. As defined in the following discussion, the best model is the one that minimizes AIC or DIC.

The procedure we followed for choosing the best model was as follows. First, we divided the data into two sets, a training set $\boldsymbol{y}_{obs}^{S1}$ and a validation set $\boldsymbol{y}_{obs}^{S2}$. (We employed a cross-validation approach to model selection.) The training and validation data sets used for Mauna Loa, Harvard Forest and the University of Utah site were defined using the dates found in Table 5.3.

Table 5.3: Definitions of training and validation data sets

	Training set (S1)		Validation set (S2)	
Site	Start	End	Start	End
Mauna Loa	1-Jan-1993	31-Dec-1995	1-Jan-1996	31-Dec-1998
Harvard Forest	1-Jan-1993	31-Dec-1995	1-Jan-1996	31-Dec-1998
Utah	1-Jan-2002	31-Dec-2004	1-Jan-2005	31-Dec-2007

Ordinary least squares parameters were estimated for each model using the training set S1. Because the dynamic leaf model (L) does not have an analytical solution the forward sensitivity methods with numerical integration of the ODEs to obtain gradients and Hessians for optimization. At this point, the mean $\hat{\boldsymbol{\theta}}$ and covariance $\hat{\boldsymbol{\Sigma}}$ of the parameters were available. The residuals were assumed to be independent and Gaussian with time-independent variance $\hat{\sigma}^2$ obtained as the sample variance of the residuals:

$$\hat{\sigma}^2 = \frac{1}{n_{obs}^{S1} - 1} \sum_{j=1}^{n_{obs}^{S1}} \left[y_{j,obs}^{S1} - \mathcal{M}(\boldsymbol{\theta}, t_j) \right]^2 \tag{5.15}$$

where n_{obs}^{S1} is the number of observations in the training set. Now everything is set for the model to be tested against the validation data. The RSS was obtained as

$$RSS = \sum_{j=1}^{n_{obs}^{S2}} \left[y_{j,obs}^{S2} - \mathcal{M}(\boldsymbol{\theta}, t_j) \right]^2 \tag{5.16}$$

Akaike's information criterion was obtained using RSS from above and computing

$$AIC = 2k + n_{obs}^{S2} \ln RSS \tag{5.17}$$

where k is the number of parameters in the model being considered and n_{obs}^{S2} is the number of observations in the validation set.

The deviance information criterion is given by

$$DIC = D(\hat{\boldsymbol{\theta}}) + 2p_D \tag{5.18}$$

where $D(\hat{\boldsymbol{\theta}})$ is the "Bayesian deviance", a measure of fit, and p_D is a measure of the effective number of parameters in the model. DIC thus represents a tradeoff between goodness-of-fit and complexity [31].

Calculating DIC is slightly more involved. It requires drawing random samples from the posterior distribution of $\boldsymbol{\theta}|\boldsymbol{y}_{obs}^{S1}$, which in our case is already determined as $N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}})$. (The actual posterior distribution may be nonnormal. We are assuming it can be approximated by a normal distribution.) We used the following steps to obtain values for DIC.

1. Choose the number of random trajectories to build. Call it n_{sim}.

2. For $i = 1..n_{sim}$:

   - Draw a random sample $\boldsymbol{\theta}^{(i)}$ from the multivariate normal distribution having mean $\hat{\boldsymbol{\theta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$.
   - Simulate the model forward in time to obtain predicted values $\boldsymbol{y}_{pred}^{S2}$ for the validation data set.
   - Calculate the sum of squared residuals for the validation data given $\boldsymbol{\theta}^{(i)}$ as

$$RSS^{(i)} = \sum_{j=1}^{n_{obs}^{S2}} \left[y_{j,obs}^{S2} - \mathcal{M}(\boldsymbol{\theta}^{(i)}, t_j) \right]^2 \tag{5.19}$$

   - Calculate the log-likelihood of the validation set's observations given the random parameter value $\boldsymbol{\theta}^{(i)}$ as

$$\ln \left[f(\boldsymbol{y}_{obs}^{S2} | \boldsymbol{\theta}^{(i)}) \right] = n_{obs}^{S2} \ln \left(\frac{1}{\sqrt{2\pi}\hat{\sigma}} \right) - \frac{RSS^{(i)}}{2\hat{\sigma}^2} \tag{5.20}$$

   - Calculate the adjusted mean deviance as

$$\begin{aligned} \bar{D} &= \frac{-2}{n_{sim}} \sum_{i=1}^{n_{sim}} \ln \left[f(\boldsymbol{y}_{obs}^{S2} | \hat{\boldsymbol{\theta}}^{(i)}) \right] \\ &= -2n_{obs}^{S2} \ln \left(\frac{1}{2\pi\hat{\sigma}^2} \right) + \frac{1}{\hat{\sigma}^2} \overline{SSR^{(i)}} \end{aligned} \tag{5.21}$$

where

$$\overline{SSR^{(i)}} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} SSR^{(i)} \tag{5.22}$$

- Calculate the adjusted deviance given the mean value $\hat{\boldsymbol{\theta}}$. This can be obtained easily since RSS was already computed earlier. We first compute the log-likelihood, which is given by

$$\ln\left[f(\boldsymbol{y}_{obs}^{S2}|\hat{\boldsymbol{\theta}})\right] = n_{obs}^{S2} \ln\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}}\right) - \frac{RSS^{(i)}}{2\hat{\sigma^2}} \tag{5.23}$$

  The deviance of the mean is then obtained by multiplying by -2, so that

$$\begin{aligned} D(\hat{\boldsymbol{\theta}}) &= -2\ln\left[f(\boldsymbol{y}_{obs}^{S2}|\hat{\boldsymbol{\theta}})\right] \\ &= -n_{obs}^{S2} \ln\left(\frac{1}{2\pi\hat{\sigma}^2}\right) + \frac{RSS}{\hat{\sigma}^2} \end{aligned} \tag{5.24}$$

- The DIC can now be obtained as

$$\begin{aligned} DIC &= 2\bar{D} - D(\hat{\boldsymbol{\theta}}) \\ &= -n_{obs}^{S2} \ln\left(\frac{1}{2\pi\hat{\sigma}^2}\right) + \frac{2\overline{RSS^{(i)}} - RSS}{\hat{\sigma}^2} \end{aligned} \tag{5.25}$$

5.3 Training and validation results

The models described above were trained to the data from Mauna Loa, Harvard Forest and Salt Lake using (nonlinear) least squares. The parameter estimates are given in Table 5.4. After training the parameters, the validation data sets defined earlier were used to rank the models. The results, which can be found in Table 5.5, support the simpler extended radiation based model (SRO) over its relatively more complex dynamic leaf model counterparts. The majority of the variance in the residuals is coming from diurnal fluctuations which neither model can capture, so it makes sense the simpler model should prevail.

Figure 5-11 illustrates the ability of Model SRO to capture seasonal variability at Mauna Loa, Harvard Forest, and University. The extended solar radiation model provides good fits to the seasonal changes in the data.

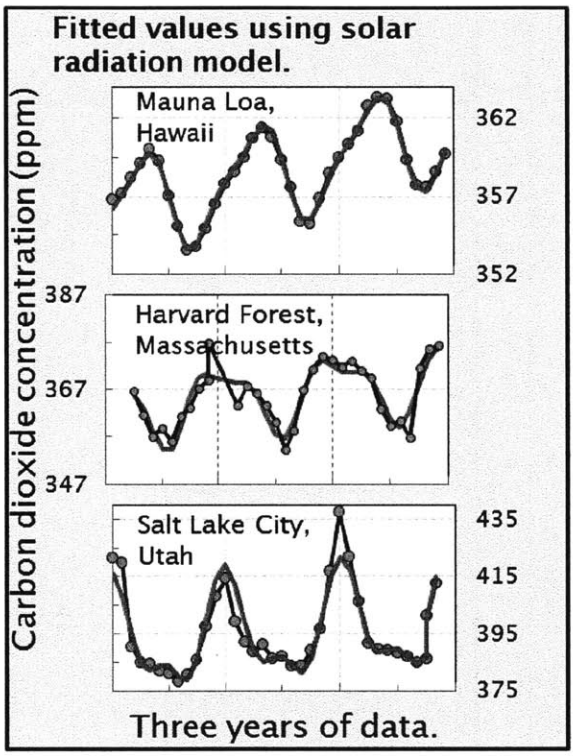

Figure 5-11: Performance of the extended solar radiation model. Nonlinear least squares fits are shown (red line) for training data from Mauna Loa, Harvard Forest and Salt Lake. Monthly averages of the original data are plotted for comparison.

Table 5.4: Least squares parameter estimates

Mauna Loa

Model	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	σ^2
S	356.7±0.03	0.0034±0.0000									5.7
E	356.7±0.03	0.0570±0.0008									5.7
SR	375.0±0.07	0.0048±0.0000	0.015±0.000	125.4±0.2							1.5
ER	375.0±0.07	0.0800±0.0004	0.015±0.000	125.4±0.2							1.5
SRO	377.0±0.05	0.0050±0.0000	0.017±0.000	115.0±0.1					-0.2±0.0	1.5±0.0	0.6
ERO	377.0±0.05	0.0842±0.0003	0.017±0.000	115.0±0.1					-0.2±0.0	1.5±0.0	0.6

Harvard Forest

Model	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	σ^2
S	360.3±0.2	0.0084±0.0003									118.1
E	360.3±0.2	0.1396±0.0046									118.1
SR	383.1±0.4	0.0060±0.0003	0.021±0.000	31.8±0.8							96.0
ER	383.2±0.4	0.0997±0.0043	0.021±0.000	31.8±0.8							96.0
SL	366.4±0.2	0.0072±0.0002	2.057±0.365	281.0±1.4	4.6±0.7	0.037±0.004	7.4±1.3	6.4±2.6			88.8
EL	366.6±0.2	0.1269±0.0041	2.073±0.398	280.7±1.3	4.6±0.7	0.035±0.004	7.4±1.5	6.9±2.7			88.9
SRO	379.5±0.4	0.0070±0.0002	0.018±0.000	33.3±0.8					-3.7±0.1	-1.2±0.1	89.2
ERO	379.5±0.4	0.1173±0.0042	0.018±0.000	33.4±0.8					-3.7±0.1	-1.2±0.1	89.2
SLO	366.5±0.2	0.0069±0.0002	2.129±0.469	280.8±1.6	4.8±0.8	0.041±0.006	7.1±1.6	5.9±2.7	-0.3±0.2	0.5±0.2	88.7
ELO	366.0±0.2	0.1262±0.0041	1.966±0.402	280.8±1.3	4.1±0.6	0.044±0.006	7.2±1.5	6.9±2.8	-0.4±0.2	0.7±0.2	88.7

Utah

Model	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	σ^2
S	393.1±0.3	0.0051±0.0005									549.1
E	393.1±0.3	0.0719±0.0066									549.1
SR	441.4±0.5	0.0079±0.0004	0.050±0.000	21.6±0.6							386.6
ER	441.5±0.5	0.1109±0.0058	0.050±0.000	21.6±0.6							386.6
SL	408.0±0.4	0.0083±0.0005	3.992±0.707	275.5±0.7	4.5±0.6	0.091±0.010	7.1±1.2	2.0±0.5			374.9
EL	407.6±0.4	0.1399±0.0057	3.585±0.580	275.6±0.7	4.7±0.5	0.087±0.007	8.0±1.3	2.6±0.7			375.4
SRO	439.3±0.5	0.0080±0.0004	0.048±0.000	23.3±0.6					0.3±0.2	6.4±0.2	366.9
ERO	439.4±0.5	0.1125±0.0057	0.048±0.000	23.3±0.6					0.3±0.2	6.4±0.2	366.8
SLO	401.7±0.3	0.0097±0.0004	4.424±0.600	281.8±2.3	7.2±0.8	0.085±0.007	8.0±1.0	7.5±3.9	7.0±0.2	11.7±0.2	378.4
ELO	400.7±0.3	0.1463±0.0057	3.986±0.767	281.4±2.1	5.5±0.7	0.086±0.008	7.8±1.5	8.5±4.4	6.6±0.2	11.4±0.2	379.7

[a] Least squares estimates are given as means ± one standard deviation. σ^2 is the sample variance of the least squares residuals from the Training data set described previously.

Table 5.5: Model validation results

Mauna Loa

Model	No. parameters	Training result (S1)	Validation results (S2)			
	k	σ	$n_{obs}^{(S2)}$	RSS	AIC	DIC
S	2	2.39	24,528	2.053E+05	3.000E+05	1.236E+05
E	2	2.39	24,528	1.974E+05	2.991E+05	1.225E+05
SR	4	1.24	24,528	4.574E+04	2.632E+05	8.546E+04
ER	4	1.24	24,528	4.458E+04	2.626E+05	8.466E+04
SRO	6	0.80	24,528	2.413E+04	2.475E+05	7.196E+04
ERO	6	0.80	24,528	2.478E+04	2.482E+05	7.317E+04

Harvard Forest

Model	No. parameters	Training result (S1)	Validation results (S2)			
	k	σ	$n_{obs}^{(S2)}$	RSS	AIC	DIC
S	2	10.87	21,630	2.949E+06	3.222E+05	1.681E+05
E	2	10.87	21,630	2.990E+06	3.225E+05	1.684E+05
SR	4	9.80	21,630	2.000E+06	3.138E+05	1.594E+05
ER	4	9.80	21,630	2.017E+06	3.140E+05	1.596E+05
SL	8	9.42	21,630	2.009E+06	3.139E+05	1.609E+05
EL	8	9.43	21,630	2.187E+06	3.158E+05	1.632E+05
SRO	6	9.44	21,630	1.894E+06	3.127E+05	1.582E+05
ERO	6	9.44	21,630	1.918E+06	3.129E+05	1.584E+05
SLO	10	9.42	21,630	1.938E+06	3.132E+05	1.607E+05
ELO	10	9.42	21,630	2.146E+06	3.154E+05	1.618E+05

Utah

Model	No. parameters	Training result (S1)	Validation results (S2)			
	k	σ	$n_{obs}^{(S2)}$	RSS	AIC	DIC
S	2	23.43	25,433	1.105E+07	4.125E+05	2.274E+05
E	2	23.43	25,433	1.111E+07	4.126E+05	2.274E+05
SR	4	19.66	25,433	8.833E+06	4.068E+05	2.211E+05
ER	4	19.66	25,433	9.022E+06	4.073E+05	2.217E+05
SL	8	19.36	25,433	1.284E+07	4.163E+05	2.155E+05
EL	8	19.38	25,433	9.015E+06	4.073E+05	2.319E+05
SRO	6	19.15	25,433	8.676E+06	4.063E+05	2.206E+05
ERO	6	19.15	25,433	8.870E+06	4.069E+05	2.211E+05
SLO	10	19.45	25,433	1.352E+07	4.176E+05	2.190E+05
ELO	10	19.49	25,433	1.044E+07	4.111E+05	2.325E+05

[a] DIC values are based on 100 random simulations. For clarification, $\hat{\sigma}$ is the standard deviation of the residuals obtained from detrending the observations in the training step.

Chapter 6

Quantifying uncertainties at Mauna Loa, Harvard and Utah

The detection problem requires us to consider shifts in future observations before any data has been collected. This requires us to make some prediction - a forecast - about the values the observations will take in the near future and then to wait and see whether the values that are eventually observed coincide with our predictions. Were instruments perfect and nature predictable, this would be a simple problem. The dilemma is that instruments are always imprecise and nature unpredictable, so that uncertainty must be entertained. This section is aimed at developing a better understanding of where this uncertainty is coming from, what sources are most important, and what can be done to reduce it.

6.1 Observational uncertainty

Observational uncertainty is due to averaging of sub-hourly trends and instrumental error. To give an idea of the scales of uncertainty contributed by natural variability versus measurement imprecision, Figure 6-1 compares the hourly sample standard deviations s^2 for Mauna Loa, Harvard Forest, and the University site in the Salt Lake CO_2 network. All three site maintain well calibrated instruments with similar precisions. Any differences are therefore expected to arise from high frequency fluctuations driven by local sources and sinks. This expectation seems to be supported, with the sample variances at Utah being larger than those Harvard, which in turn are larger than those at Mauna Loa. In the pristine conditions of Mauna Loa, the sub-hourly samples typically vary about the mean with a standard deviation of ~0.04-1.16 ppm (with these number being the 2.5 and 97.5 percentiles of the standard deviations). The mean of the standard deviations from Mauna Loa was 0.28 ppm. In contrast, the the same percentiles of the standard deviatons from the University of Utah represent a range of 0.17-13.39 ppm, with a mean of 2.81 ppm. The bottom plot in Figure 6-1 suggests that the samples appear to be more variable in the winter than in the summer at the University of Utah. The reason this occurs is most likely due to the fact that atmospheric mixing tends to be more vigorous in the summertime because

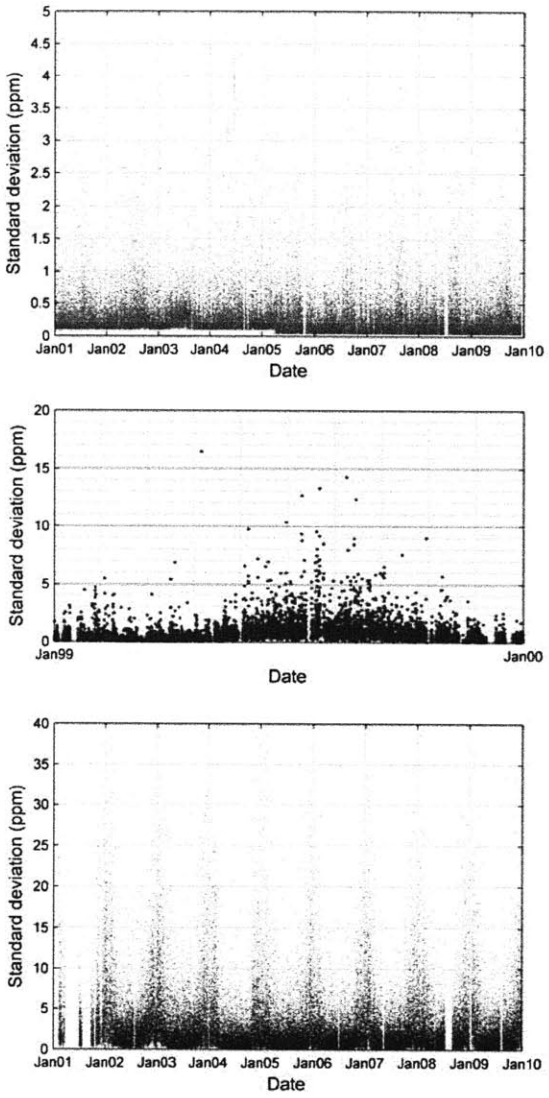

Figure 6-1: Natural variability versus measurement imprecision: contributions to observational uncertainty. Shown are hourly sample standard deviations of sub-hourly CO_2 measurements from Mauna Loa (top), Harvard Forest (middle), and the University of Utah (bottom). Atmospheric CO_2 measurements were typically made 4 times per hour at Mauna Loa by NOAA and 12 times per hour at Harvard Forest and the University of Utah. At Mauna Loa, NOAA reported the standard deviations associated with the hourly means. These were converted to sample standard deviations by multiplying by the square root of the number of samples considered per hour, which in this case was a factor of $\sqrt{4} = 2$. The other two sites report sample standard deviations of the sub-hourly observations. The connection between the standard deviation of an hourly mean value and the sample standard deviation of the sub-hourly measurements used to compute the mean can be understood by taking the square roots of the left and right sides of (2.7). Special thanks go to Bill Munger, who made the standard deviation data from Harvard Forest available.

of sunlight driven thermal instabilities in the air near the surface. The greater degree of sub-hourly variability in the wintertime could come from two places. First, because the boundary layer tends to be relatively stable, CO_2 concentrations may be changing smoothly from hour to hour. Second, plumes emitted by local sources will tend to remain more coherent in the atmosphere and will create shifts in the signal being measured as they are advected across the sensor. Table 6.1 gives statistics for the sample standard deviations from Mauna Loa and University of Utah. We calculated wintertime (October-March) and summertime (April-September) statistics separately to determine if seasonality appears to be significant at the two sites. To appreciate

Table 6.1: Statistics for sample standard deviations of sub-hourly CO_2 observations

Mauna Loa			
Season	2.5^{th} Percentile	Mean	97.5^{th} Percentile
Global	0.04	0.28	1.16
Summer	0.04	0.30	1.26
Winter	0.04	0.25	1.04

Harvard Forest			
Season	2.5^{th} Percentile	Mean	97.5^{th} Percentile
Global	0.06	0.78	3.44
Summer	0.10	1.05	4.28
Winter	0.05	0.46	2.02

University of Utah			
Season	2.5^{th} Percentile	Mean	97.5^{th} Percentile
Global	0.17	2.81	13.4
Summer	0.15	1.84	7.19
Winter	0.23	3.75	17.1

the effect location has on observational uncertainty, Figure 6-2 shows how the standard deviations associated with the hourly mean values at Mauna Loa and University of Utah evolve as the number of sub-hourly measurements being averaged grows. For each site, two curves are shown. One curve represents the standard deviations obtained if summertime measurements are considered. The second gives the same for wintertime measurements. All of the curves have been calculated by dividing the mean values in Table 6.1 by the square root of the number of samples being averaged (n_{avg}). The basis for doing this can be seen by taking the square roots of the left and right sides of equation (2.7). Figure 6-2 illustrates a few points. First, note that by sampling more frequently the uncertainty associated with the observation process can be reduced. Second, at a site like University of Utah, improving the precision of the sensor is unlikely to make much of a difference since we presume the majority of the uncertainty is coming from variability of atmospheric CO_2 crossing the sensor during

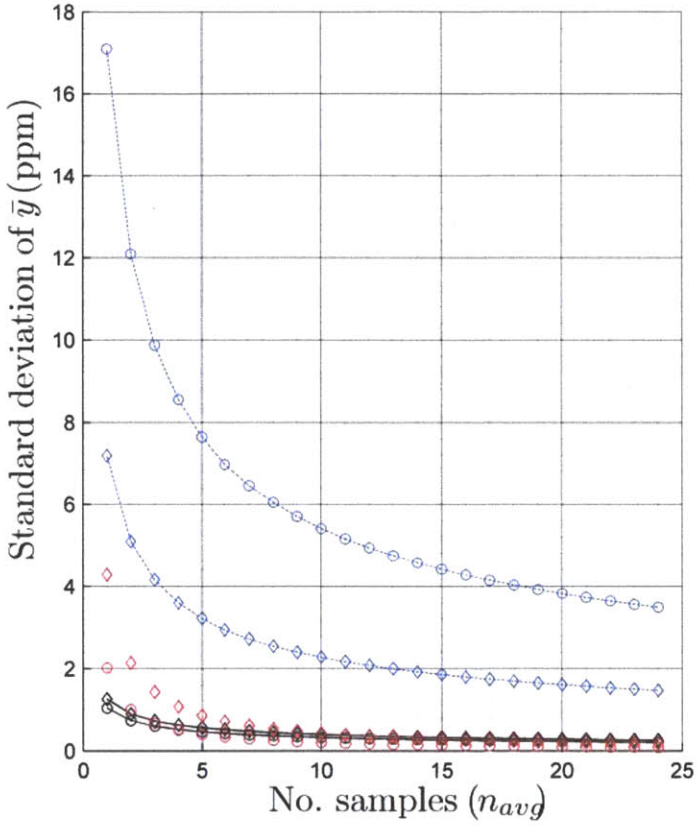

Figure 6-2: Evolution of uncertainty associated with hourly mean CO_2 at Mauna Loa, Harvard Forest, and University of Utah. Six curves are shown, two for each site. Curves for Mauna Loa are shown as solid black lines, for Harvard Forest as dotted green lines, and those for University of Utah as dashed blue lines. For each site, the summertime results are marked by diamonds and the wintertime results are marked by open circles. Each curve has been calculated as the 97.5^{th} percentile value for the respective site and season tabulated in Table 6.1 divided by the square root of the number of samples recorded per hour (n_{avg}). The reason we used the 97.5^{th} percentile instead of the mean value was to capture the extreme cases where the sub-hourly samples are most variable. To give an idea of the measurement frequencies in place at actual sites, Mauna Loa typically records 4 samples per hour and the other sites typically record 12 per hour. The curves shown measure the uncertainty associated with the hourly means. They can be interpreted as follows. Suppose samples are recorded at a rate of 12 per hour at a pristine location like Mauna Loa. The standard deviation of the hourly means will be less than roughly 0.4 ppm about 97.5 percent of the time. The standard error of the hourly mean is 1.96 times this value, or roughly 0.8 ppm. Therefore, if the sample mean from a given hour is 390 ppm, the true mean value from that hour should lie in the confidence interval 390±0.8 ppm. Sampling at the same rate but in a dirty location like the University of Utah the uncertainty we might expect the spread of the confidence interval to depend on the season. In the summer, the sample standard deviation is roughly 2 ppm, which translates into a confidence interval of approximately 390±4 ppm. In the winter, the standard deviation is roughly 5 ppm, which translates into a confidence interval of approximately 390±10 ppm.

the course of a given hour and not from imprecision or drift of the instrumental calibration. If this sub-hourly varaibility is random with respect to time, then sampling more times per hour should reduce the uncertainty associated with the hourly mean. If the variability is systematic, however, then sampling more frequently will lead to overoptimistic statements about the variance of the hourly mean and a better way to proceed would be to fit the sub-hourly measurements to a line and use the variance of the straight-line prediction as the variance of the hourly mean. At a pristine site like Mauna Loa, it is still unclear what fraction of the variability is due to the instrument, since we have considered measurements the site curators usually reject because of the influence of local sources and sinks. For all of these reasons, it is difficult to come to a strong conclusion about how to reduce observational uncertainty for any of the sites under consideration. The best we can do is to use values from the curves in Figure 6-2 and hope that sub-hourly variability is mostly random so that the values obtained are representative of the actual uncertainty in the hourly means from those sites.

6.2 Misfit uncertainty

The contribution from model misfit is captured in part by the residuals. Part of the variability in the residuals is due to observation error, since the model errors are given by

$$\underbrace{\varepsilon(t)}_{\text{Model error}} = \underbrace{\varepsilon_{\mathcal{M}}(t)}_{\text{Misfit error}} + \varepsilon_{obs}(t) \tag{6.1}$$

As we will see, observation error is negligible compared to the residuals at Harvard and Utah, so the variance of the residuals can be used to approximate the variance of misfit errors at these sites. At Mauna Loa the same is not true, but to keep things simple we use residual variance as a surrogate for misfit error.

Table 6.2 lists the standard deviations of the least squares residuals obtained using models S and SRO at Mauna Loa, Harvard Forest and the University of Utah. We expect two trends to be present in the standard deviations. First, as diurnal fluctuations in the hourly means increase in magnitude - that is, as we move from pristine conditions at Mauna Loa to "dirtier" conditions at Harvard Forest and Utah - the residuals should tend to increase, implying greater misfit uncertainty. Second, as the model is improved by adding a seasonal predictive capability the residuals should decrease. We expect the gain from improved modeling to be less significant at Utah, where the variability is presumably dominated by urban emissions and not biology, than at Mauna Loa and Harvard Forest, for which the seasonal component is relatively more important.

Table 6.2: Scales of misfit uncertainty

	Residual standard deviation (ppm)[a]	
Site	Interannual[b]	Seasonal[b]
Mauna Loa	2.39	0.80
Harvard Forest	10.87	9.44
University of Utah	23.43	19.15

[a] The residuals were obtained by subtracting the least squares best fit trajectory for models S (interannual) and SRO (seasonal) from the training data.

We notice a few things in Table 6.1. First, the diurnal fluctuations at Mauna Loa are presumably small (they would probably be smaller yet if we rejected samples affected by local sources and sinks), so the majority of the 2.39 ppm standard deviation of the observations around the linear trend is most likely coming from seasonality. This is qualitatively confirmed by the large reduction in the residuals after seasonality is accounted for by model SRO. At Harvard Forest, seasonality is strong but we expect diurnal fluctuations to be stronger. Still, the leaf area model may not be doing a good job fitting the seasonal trend. To test this, we generated the power spectra of the residuals from the linear and seasonal models already described. The results are shown in Figure 6-3. Figure 6-3 tells us a few things. First, the observations exhibit strong annual and diurnal signals. Second, the residuals are still seasonal after removal of a linear trend. Third, it suggests that at least qualitatively the majority of the seasonality is removed by adding the solar radiation and oceanic terms. The surprising thing is that the standard deviation of the residuals only decreases from 10.87 to 9.44 ppm. Even after removing the seasonal trend, the residuals are still much more variable at Harvard Forest than at Mauna Loa. Based on the power spectra in Figure 6-3, the logical explanation for this is the strong diurnal fluctuations present at Harvard Forest, which we expect are largely absent from the Mauna Loa data. Therefore, we suspect **diurnal variability at dirty sites like Harvard Forest and Utah represents a significant contribution to misfit uncertainty when the model only accounts for seasonality and we might think about reducing misfit uncertainty by somehow removing diurnal fluctuations before testing for a shift**. There are different ways to reduce diurnal variability. We can improve the model, for example by adding correlative terms linking fluctuations in CO_2 with fluctuations in carbon monoxide and carbon and oxygen isotopes. An analogous approach is to use signals from highly correlated, upwind sensors to explain fluctuations in readings from the sensor of interest. Finally, if all else fails one can apply a heuristic filter like a simple moving average. Each of these possible alternatives are considered next.

Smoothing the observations with a T-day moving average.
Here we consider the use of simple moving averages to reduce the diurnal signal that is present at sites like Harvard Forest and Utah. A T-day moving average entails

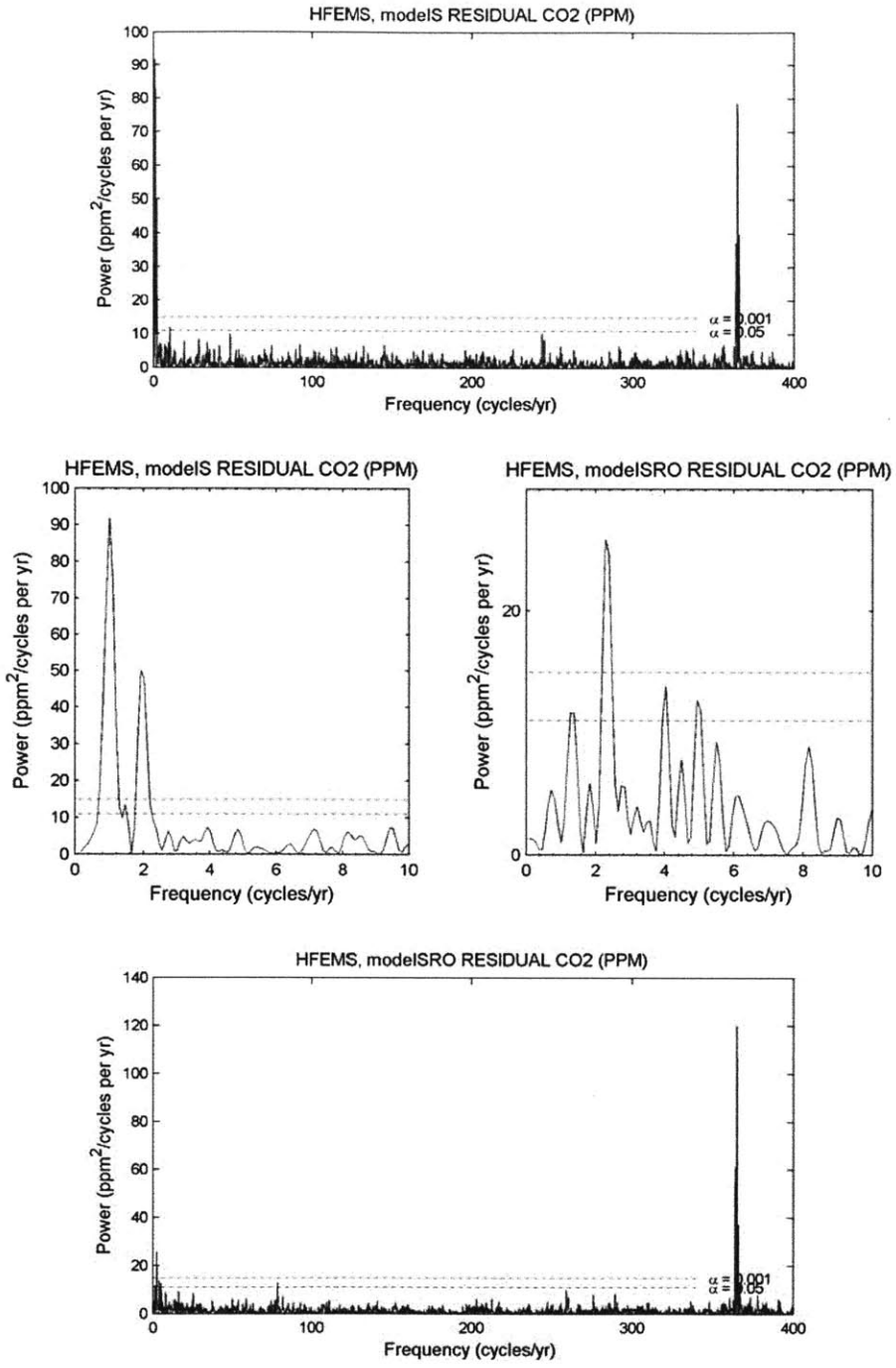

Figure 6-3: The impact of modeling seasonality on the harmonic content of the residuals. Power spectra of residuals obtained after subtracting interannual (top and middle left, model S) and leaf-model based seasonal (bottom and middle right, model SRO) trends from the Harvard Forest training data. The plots in the middle zoom in on the low frequency signals. Initially, both annual and a semiannual signals are present. Model SRO removes the annual signal and most of the semiannual signal from the observations. The power spectra were determined using 10^3 random samples from the least squares residuals. Because the data are unevenly spaced in time, the power spectra were calculated using Lomb's method. The dashed lines mark the power thresholds asociated with 0.05 and 0.001 significance levels.

forming a window containing samples within $\pm T/2$ days to either side of the time of interest and then averaging the samples in the window using an equation like 2.2. Figure 6-4 illustrates the effect of applying such a filter to hourly means from Harvard Forest, which is to reduce the variability of the hourly values about the modeled trend. The downside is that in situ CO_2 series are plagued by frequent gaps, so that the number of representative smoothed values tends to decrease with increasing window size. Figure 6-5 shows the effect of applying a one day moving average to hourly mean

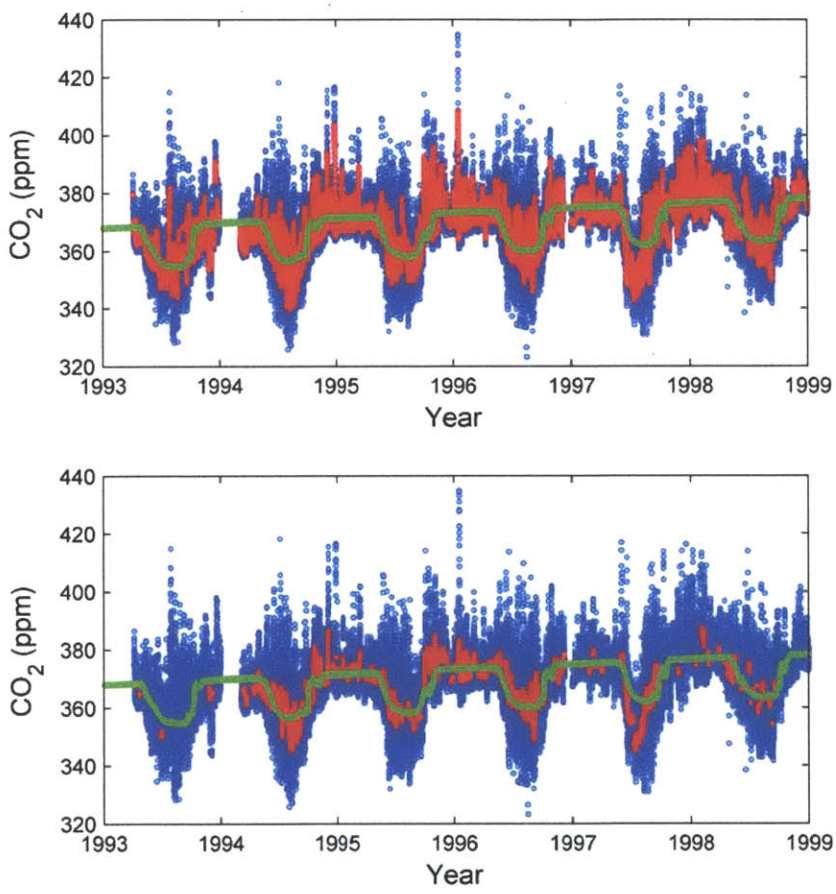

Figure 6-4: Applying moving average filters smoothens the hourly observations from Harvard Forest. The original hourly means are shown in blue, the hourly filtered values in red, and the green line is the best fit from model SL employing the dynamic leaf model of Liu et. al [42]. Filtered values have been obtained by applying T-day moving windows about the hours of interest. Top: 1-day moving average. Bottom: 4-day moving average. Only moving averages for which at least 50% of the hourly means were available are plotted.

CO_2 observations from Harvard Forest. As we can see, the diurnal signal is eliminated to the level of the noise. Time series methods like moving averages are not the most ideal way to reduce the diurnal signal. The reason for this is that they are devoid of any physics. While the correlative models will no doubt tend to oversimplify the processes contributing to diurnal fluctuations in CO_2, they do have a physical basis and can take in real data to inform future expectations. Nonetheless, it is still

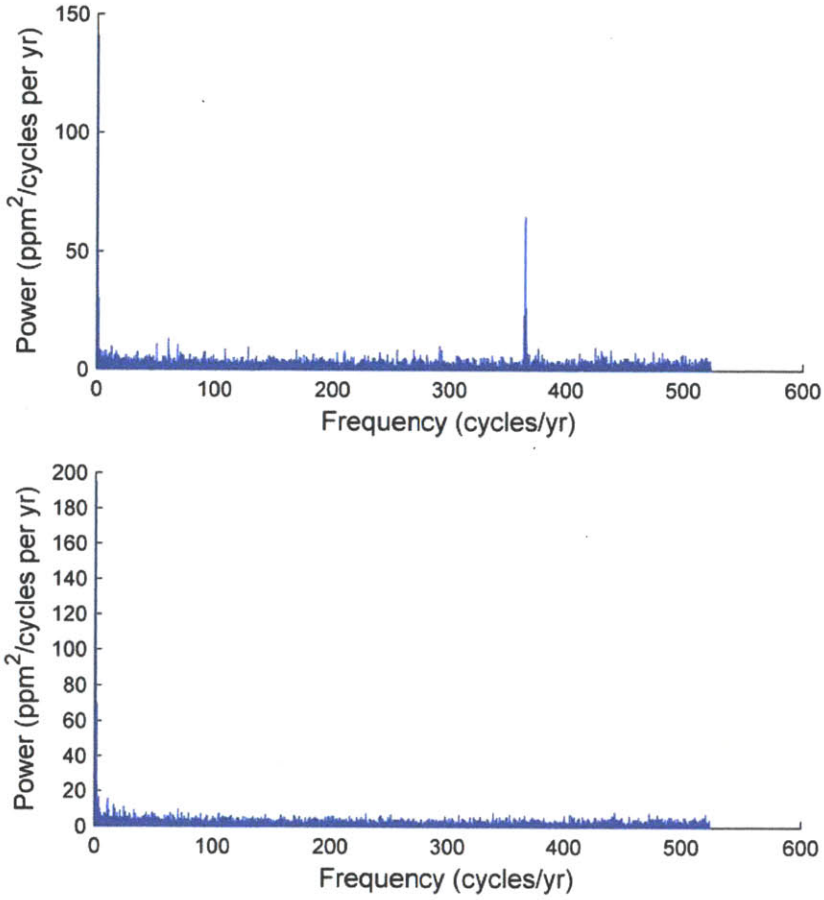

Figure 6-5: Smoothing eliminates the diurnal signal. Two power spectra are shown. Top: the power spectrum corresponding to the original hourly mean CO_2 observations from Harvard Forest. Bottom: the power spectrum after smoothing the original observations with a 1-day moving average filter. Lomb periodograms were computed because of uneven spacing of the data in time. Six years of data were considered including dates considered were Jan 1, 1993 through Dec 31, 1998.

interesting to see what impact smoothing has on forecast uncertainty. Smoothing affects two of the three error terms contributing to the forecast: observational and misfit errors. We expect observational uncertainty to increase because, as before with the case of hourly averages, representing the true signal as a T-day mean by averaging implies error due to the presence of the systematic diurnal signal the average is meant to eliminate. Averaging is not all bad though. The advantage is the reason why we are thinking about averaging in the first place, which is to reduce the misfit between the observations and seasonal model trajectories (as illustrated in Figure 6-4.)

Table 6.3: Effect of smoothing on observational uncertainty

	Stdev[a](ppm)		
T (days)	Global[b]	Summer[b]	Winter[b]
0[c]	0.99	1.24	0.58
1	2.95	3.35	1.65
2	2.02	2.21	1.22
4	1.40	1.48	0.93
7	1.03	1.08	0.70

[a] T-day moving average filter was applied to the data. Only windows with 50% of original hourly means available were considered. Numbers are based on observations between Jan 1993 and Dec 1998 from Harvard Forest. Standard deviations determined as 97.5^{th} percentile of the T-day window sample standard deviations, divided by the square root of the number of hourly observations being considered (which is $24 \cdot T + 1$).

[b] Global = both summer and winter, Summer = Apr-Sep, and Winter = Oct-Mar.

[c] Standard deviations represent uncertainty in unsmoothed hourly mean observations. Values are representative ones obtained by dividing the appropriate standard deviations from Table 6.1 by $\sqrt{12}$ since Harvard Forest typically measures 12 samples per hour.

Table 6.4: Effect of smoothing on misfit uncertainty

	Stdev[a](ppm)		
T (days)	Global[b]	Summer[b]	Winter[b]
0	9.44	10.43	7.73
1	9.13	10.39	6.86
2	8.79	9.77	7.03
4	9.17	10.04	7.44
7	9.79	10.62	7.98

[a] T-day moving averages were applied to the hourly mean CO_2 observations in the Harvard Forest traiing data set. The same data was used to train model SRO. Standard deviations were determined as the square roots of sample variances of the smoothed hourly values minus the least squares fit of SRO to the original hourly means. Distinct subsets of the resulting difference series were used to generate standard deviations to better appreciate the impact of smoothing on misfit in different seasons (Summer, Winter or Global which is both Summer and Winter). Only windows with 50% of original hourly means available were considered.

[b] Global = both summer and winter, Summer = Apr-Sep, and Winter = Oct-Mar.

Correlating diurnal fluctuations with upwind CO_2 series

Figure 6-6 below compares CO_2 observed at Residential and Junior High sites in Salt Lake to the University site. The dominant wind direction is from the north and the sites are roughly aligned from north to south. University is the northernmost site, with Residential being about 5 km to the south and Junior High about 10 km to the south. We therefore expect concentrations at Residential and Junior High to be correlated with some lag time related to the time required for advection. Because Junior High is farther from University, we also expect the diurnal fluctuations present in its observations to be less correlated with University's than those at Residential. Several points are highlighted by the figure. First, note that the seasonal fitted trends line up with what we expect. That is, the fitted values at Residential and Junior High are greater than those for University, reflecting the fact that air traveling from University to the other sites encounters urban emissions along the way. Second, note that the afternoon minima tend to align quite well - both in time and in magnitude - probably due to enhanced vertical mixing. Third, in the morning and at night, the air is CO_2 rich. Fourth, during the window of time considered in the figure, the nighttime peaks in CO_2 tend to be higher at Residential and Junior High than at University. This relates back to the first point. Fifth, the timing of the diurnal increase and decline appears to be similar at the three sites. Finally, note that the nighttime peaks at Junior High, which is the southernmost point, tend to be much higher than those at Residential. There are several nights however where the nighttime peak at Junior High is visibly closer to that at University than the peak at Residential is. This reason for this is somewhat of a mystery.

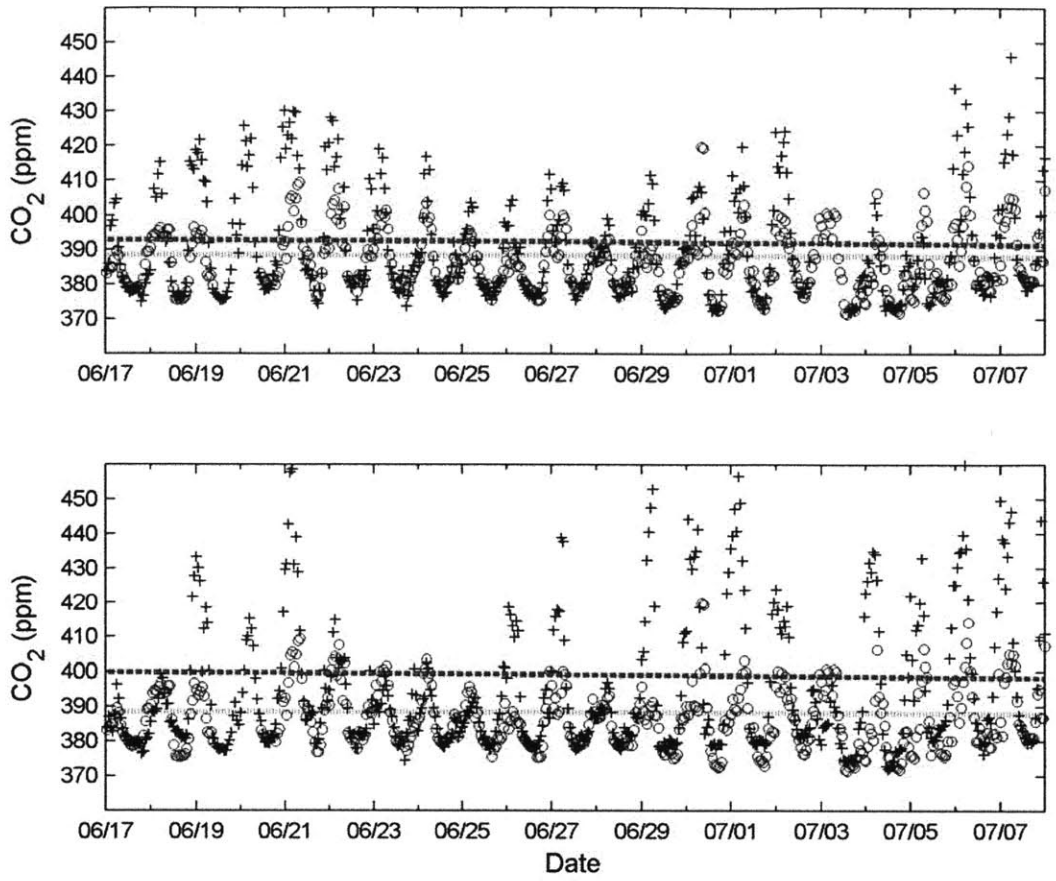

Figure 6-6: Comparing CO_2 series along a 10 km north-south transect of Salt Lake City, Utah. Top: Hourly mean CO_2 observations from University (open blue circles) and Residential (black crosses) sites. The least squares trajectories obtained by fitting model SRO to a training set consisting of dates from Jan 1, 2005 through Dec 31, 2007 are shown in red, with University being the dotted line and Residential the dashed line. Bottom: same as in the top figure, but with data from Junior High replacing the data from Residential. Dates shown in both plots are 00:00:00 Jun 17 to 00:00:00 Jul 8, 2005.

To improve our ability to detect, the ideal neighboring sensor(s) must exhibit diurnal fluctuations that are highly correlated with the ones at the location of interest but should also be devoid of the leak signal. The time series in Figure 6-6 often exhibit similar timing. However, a simple correlation model will fail to predict the magnitude of the difference between two series like the ones from University and Residential. It is possible that atmospheric mixing plays a role, but there may also be times when a highly local source like a stationary automobile leads to large divergences from the CO_2 baseline. In such cases, measurements such as carbon monoxide concentration at the location of interest could help differentiate automobile emissions from emissions due to leakage by accounting for the fact that automobile emissions will tend to increase CO concentration as well as CO_2 concentration. Meteorological measurements such as the Monin-Obhukov stability parameter and wind direction may prove useful for explaining changes associated with boundary layer dynamics and advective

transport. Finally, we note that when considering placing neighbor sensors in the field, one should account for the natural correlation structure of atmospheric CO_2. In desiging a network to detect a leak from a geological storage site, a second length scale also becomes important, which is the expected footprint of a leak. Leakage over a large area could mean a sensor meant to act as an upwind sensor exhibits the leak signal just like the candidate sensor, so that in subtracting the neighbor-based trend we may remove some or all of the leak signal we are trying to detect. This spatial aspect of the network design problem merits further attention.

6.3 Forward uncertainty

Forward uncertainty arises when uncertain inputs and model parameters are passed through a model to obtain predictions. Figure 6-7 illustrates the impact parametric uncertainty has on model predictions. Because the parameters we consider are time invariant, the figure was generated by taking random samples from the posterior distribution obtained by training the models using the Training data set to obtain least squares values. The posterior distributions are multivariate normal densities with mean $\hat{\boldsymbol{\theta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ obtained by ordinary least squares estimation. It is apparent from the figure that the scale of the forward variability is far less than that due to model misfit for the forecast window considered. Longer forecasts will entail more uncertainty, which is readily apparent from the standard deviation plots in the figure. Additionally, the forecast uncertainty decreases as we add information to the model.

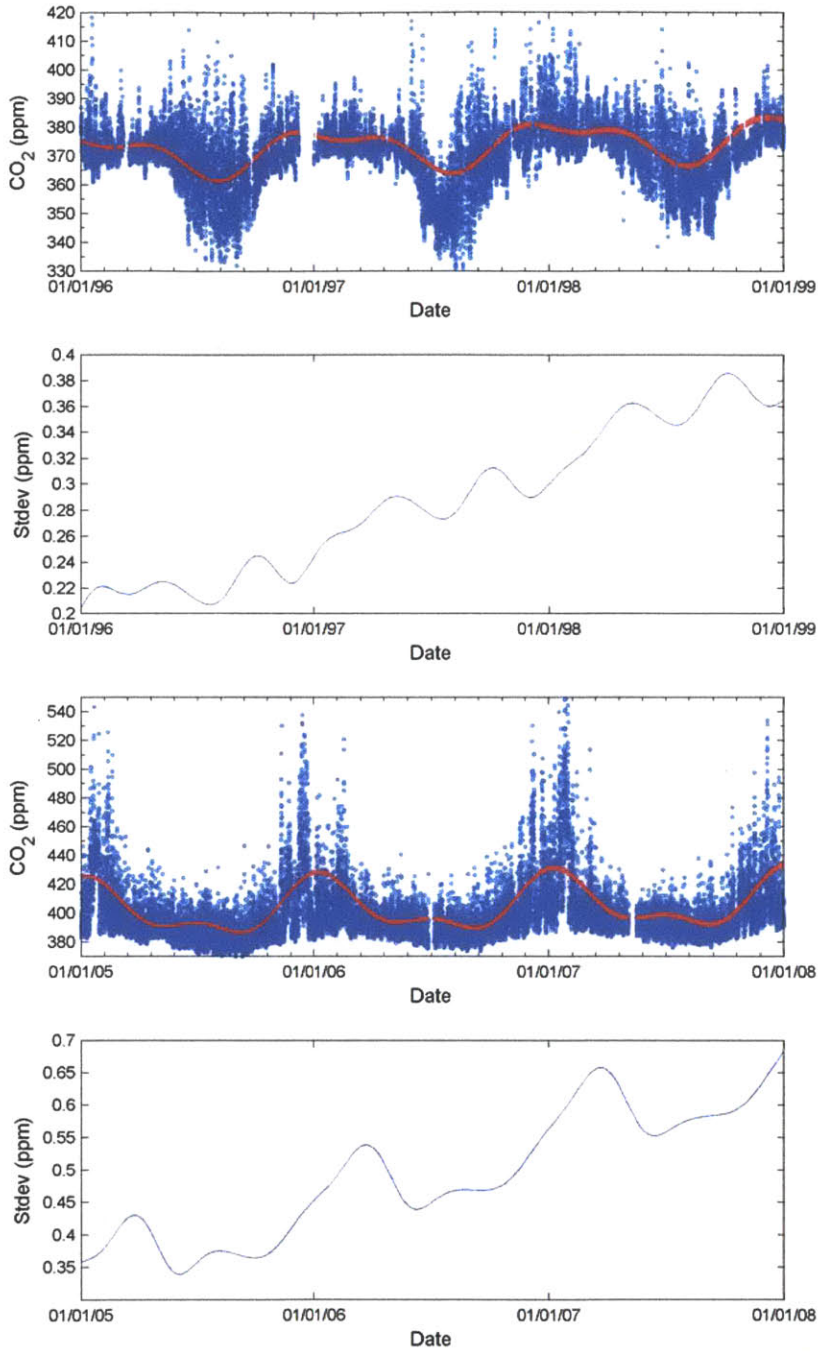

Figure 6-7: Uncertainty in forward simulations. From the top as follows. First and third axes plot 100 random forward model SRO trajectories (red curve) alongside hourly means (blue circles) for Harvard Forest and University of Utah, respectively. Second and fourth axes plot the evolution of the standard deviations of the random trajectories in time. Each random trajectory is based on a random draw from the multivariate normal posterior density defined by the least squares estimate and covariance matrix obtained using the usual Training data set.

6.4 Forecast uncertainty

Figure 6-8 is a stacked plot illustrating the effect of model, location, and smoothing on the variance associateed with our forecasts. It is readily apparent that forecast and observational uncertainty are negligible. The biggest concern is by far the uncertainty due to misfit, which in this case arises because of diurnal fluctuations that cannot be captured using sesasonal trends.

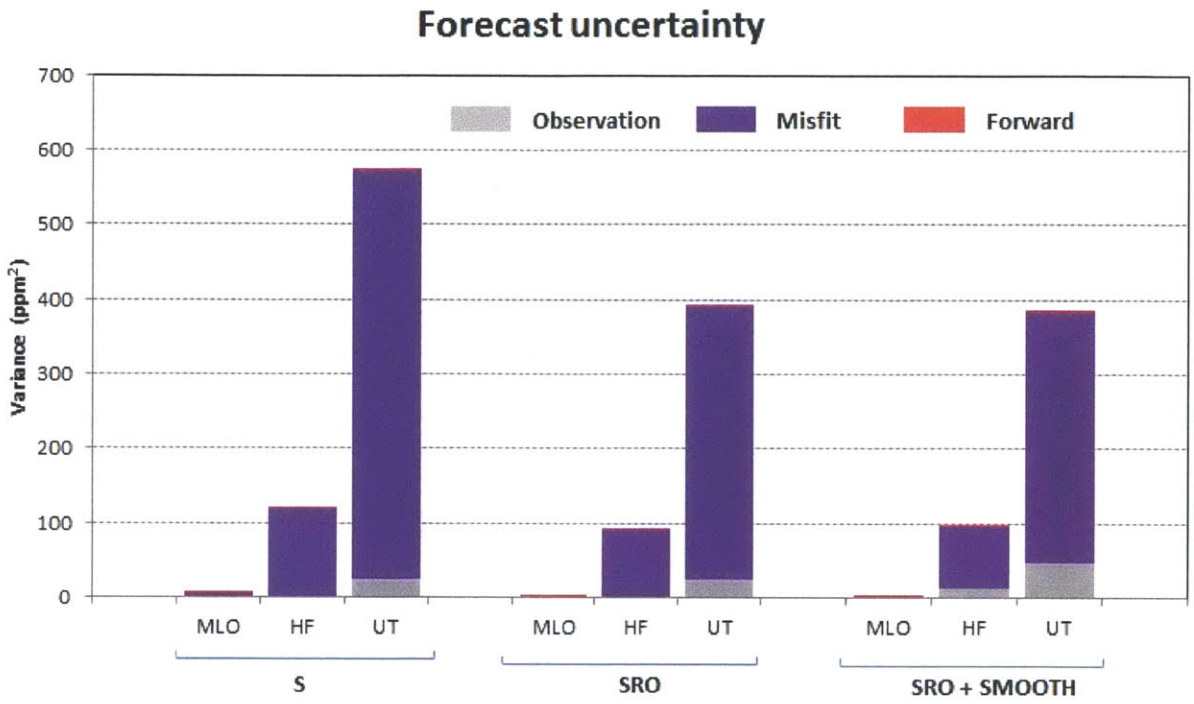

Figure 6-8: Forecast uncertainty. Shown are variances associated with forecasts. The overall forecast variance is broken into its individual contributors: observation, misfit and forward variances. Nine cases are considered. These are labeled along the horizontal axis. The nine cases are broken down into three subsets, one for model S, one for model SRO, and one for model SRO plus smoothed (1-day moving average) observations. These cases illustrate the effect of improving the model and smoothing the hourly observations to eliminate diurnal trends. In each of these cases, we determined variances for observations, misfit, and forward uncertainties. These were determined as follows. Observational variances were determined in the manner described earlier on a site-wise basis and assuming 12 samples per hour. Misfit variance was determined as the sample variance of the least squares residuals obtained by fitting the appropriate Training data set. Finally, forward variance was determined as the maximum value attained by the standard deviation of the random trajectories in the Validation data set.

6.5 Conclusions

The seasonal models' inability to capture diurnal variability in the observations is the biggest contributor to the overall uncertainty in future atmospheric CO_2 levels. Smoothing the observations by applying moving average filters with widths of at least

$T = 1$ day can help reduce misfit errors. The biggest downside to using simple filters is that they cannot account for changes in the physical system. A better approach would be to improve the model by adding correlative terms linking diurnal fluctuations in CO_2 at the location of interest to changes in CO_2 at upwind sensors as well as to ancillary variables such as CO concentration, carbon and oxygen isotopic abundances, wind direction, atmospheric mixing (via the Monin-Obhukov length), and others. A worthwhile next step would be to screen new candidate model terms using model selection criteria like RSS, AIC and DIC, which were presented earlier. In deciding where to place sensors in the field, the correlation structure of atmospheric CO_2 should be accounted for as well as the expected leak footprint. This spatial aspect of the network design problem merits further attention.

In conclusion, improving our ability to detect a shift in atmospheric CO_2 observations from a given sensor will require methods for coping with diurnal variability. The spatial aspect of the design problem merits attention.

Chapter 7

Detecting urban weekend weekday shifts

One of the challenges facing future monitoring sites will be demonstrating the ability to detect leaks before any leaks have occurred. This chapter illustrates how this can be done using the urban weekend weekday effect on CO_2 as an analogue for leakage from belowground.

7.1 Urban weekend weekday effects on CO_2

The hypothesis is that during the workweek, increased human activity leads to greater emissions of CO_2 than on the weekend. If this hypothesis is correct, then atmospheric CO_2 concentrations should be higher on weekdays than on weekends. This difference can be tested for in the absence of shifts associated with leaks.

7.1.1 Differences observed in Salt Lake City, Utah

Figure 7-1 compares weekend and weekday diurnal trends in atmospheric CO_2 at the University site in Salt Lake City. An upward shift is discernible in the morning. This is to be expected since in the morning rush hour often coincides with a stable atmospheric boundary layer.

The idea used in this chapter is to look for shifts in the distribution of the data. This is illustrated in Figure 7-2, which plots interquartile ranges for the data instead of mean confidence intervals. The challenge is distinguishing shifts given the high degree of overlap often present between distributions on weekdays and weekends.

Figure 7-3 depicts interquartile ranges by month of the year. Upward shifts are generally more noticeable in the morning. Large shifts are observed in some summer months and winter months. Larger shifts are expected in the winter, when sunlight is weak and the atmosphere relatively stable. The large shifts in the summer might reflect biological respiration present to a lesser degree in the winter.

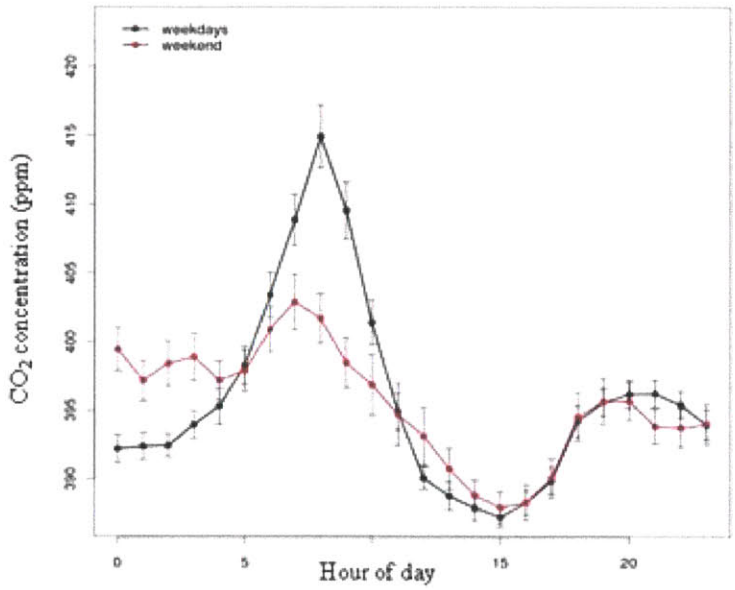

Figure 7-1: Comparison of mean diurnal atmospheric CO_2 trajectories on weekends (red) and weekdays (blue) during the month of October, 2006, at the University site in Salt Lake City, Utah. Sample means and 95% errors bars are given by hour of day. Error bars associated with the sample mean have been computed using 2.28 at the α =0.05 level of significance. This figure was adapted from McKain and Wofsy [88].

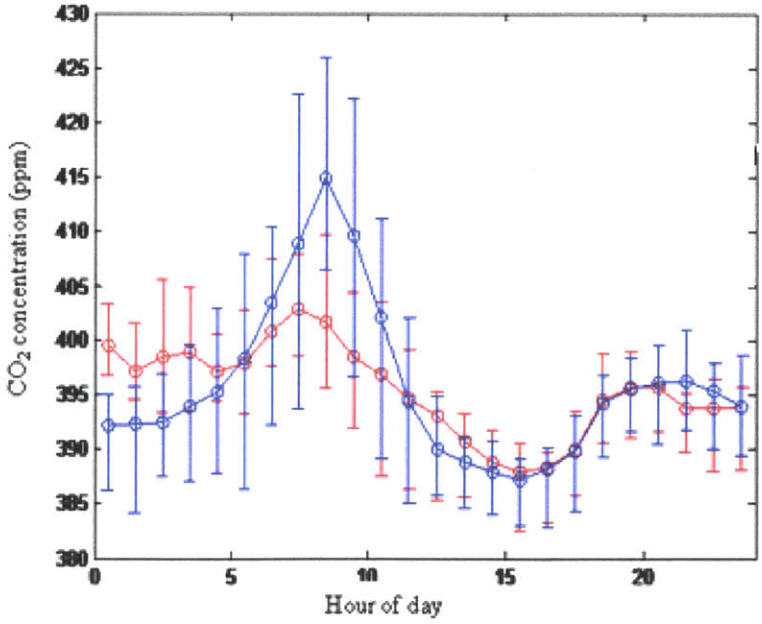

Figure 7-2: Comparison of interquartile ranges of diurnal atmospheric CO_2 trajectories on weekends (red) and weekdays (blue) during the month of October, 2006, at the University site in Salt Lake City, Utah. The interquartile ranges represent the 25^{th}-75^{th} percentiles of the original hourly averages. Conceptually, this figure illustrates a shift in the distribution of atmospheric CO_2 from weekends to weekdays for each hour of the day.

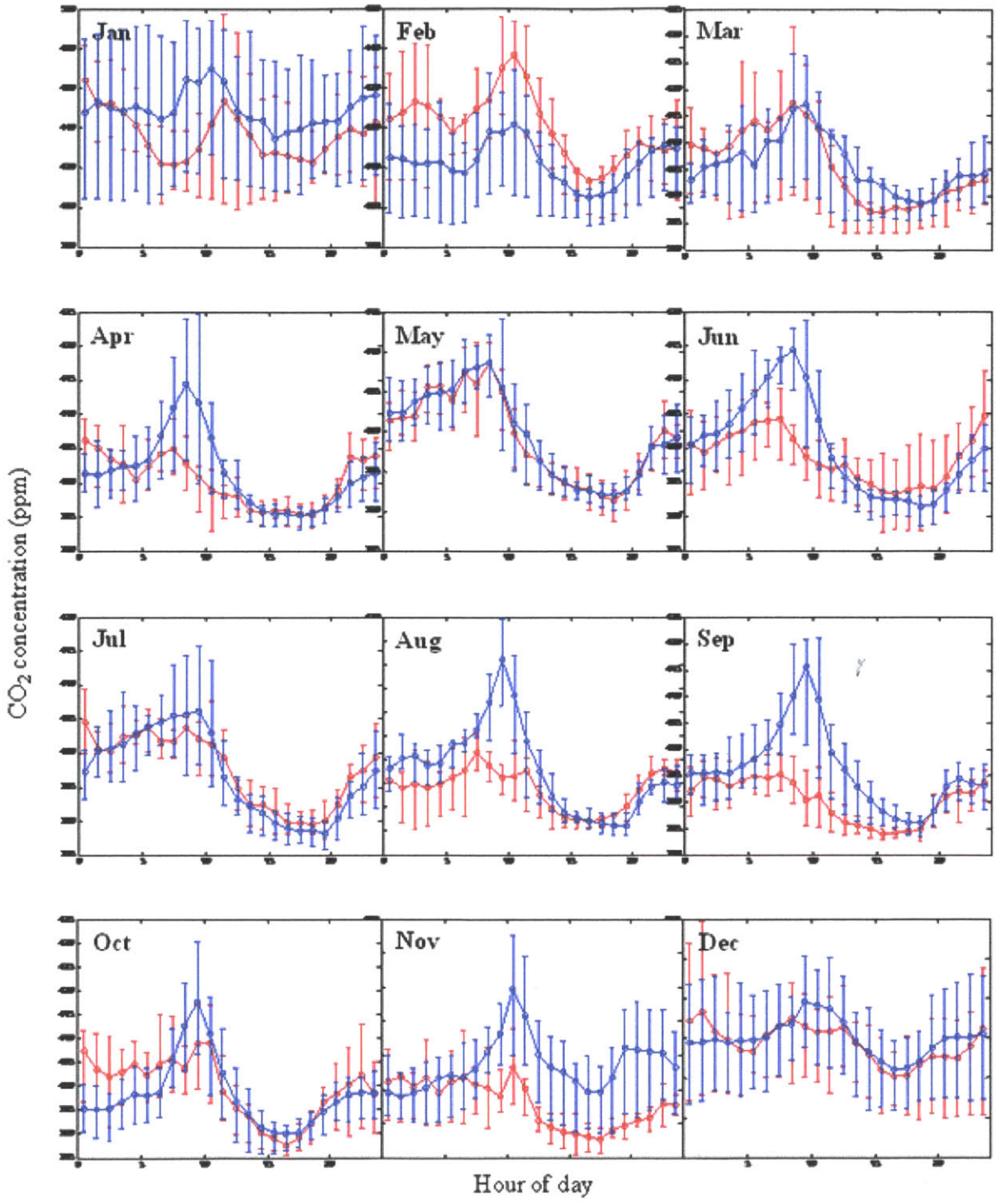

Figure 7-3: Changes in weekend weekday diurnal CO_2 trends with month of year. Interquartile ranges are shown for weekend (red) and weekday (blue) data by month and hour of day for the University site in Salt Lake City, UT, during year 2009. Interquartile ranges were constructed as described in the caption of Figure 7-2.

7.1.2 Basic analysis of weekend weekday differences at Mauna Loa, Harvard Forest and Salt Lake

Table 7.1 explores the effects of season and time of day on the size of the shift present in atmospheric CO_2 through statistics for weekend and weekday CO_2 residuals[1] at Mauna Loa, Harvard Forest and the University of Utah. Biological effects are expected to be less important in the winter (for example, see the respiration term in Figure 4-11). Weekend weekday CO_2 shifts have been observed to be bigger in the morning than the afternoon [91, 90]. Differences due to timing and season are illustrated by statistics for intersections of the Winter/Summer and AM/PM sets.

The last column in Table 7.1 gives the signal to noise ratio of the shift, as described in (3.3). The signal to noise ratio measures the ease with which the weekend weekday shift in atmospheric CO_2 should be detectable.

The signal to noise ratios of the weekend weekday shift is generally small at both Mauna Loa and Harvard Forest. At Harvard Forest, it increases to $\sim$0.15 in the winter. At the University site, the signal to noise ratio is relatively large throughout the year. The magnitudes of the signal to noise ratios tend to be larger at that site than the other two. We expect the weekend weekday shift to be more easily detected at this site than the other two. This makes sense considering University is located in a city, whereas the other two sites are relatively far from urban effects.

The samples of most interest are those in the "Winter AM" subset. Because the samples come from cold months, biological effects are expected to be small. The interest in the morning rather than afternoon reflects the expectation that morning rush hour coupled with poor mixing should lead to high levels of CO_2. This expectation seems reasonable, given that Winter Mon-Fri residuals have a mean of 7.05 ± 0.479 ppm in the morning versus -10.39 ± 0.328 in the afternoon[2].

The Winter AM shift at Utah is +7.86 ppm[3]. The 95% confidence intervals[4] for the weekend and weekday residuals are -0.81 ± 1.35 ppm and 7.05 ± 0.96 ppm, respectively. A positive weekend weekday shift is clearly present in Salt Lake City, Utah.

The Winter AM shift at Harvard Forest is +1.13 ppm. The 95% confidence intervals for the weekend and weekday residuals are -1.59 ± 0.32 ppm and -0.46 ± 0.22 ppm, respectively, which leads us to believe a positive weekend weekday shift may also be present at Harvard Forest.

The Winter AM shift at Mauna Loa is +0.019 ppm. The 95% confidence intervals for the weekend and weekday residuals are -0.027 ± 0.036 ppm and -0.008 ± 0.024 ppm. The weekend weekday shift is not statistically significant, as expected given

[1]The residuals have been obtained by detrending using model SRO. The method traditionally employed for this sort of problem is to compute confidence intervals for sample means and detect a shift when they do not overlap. (For example, this is the method presumably used by Idso et al. [89].) The approach in this chapter is conceptually similar to the one presented by Cerveny and Coakley [90], which is to consider shifts from a gradual seasonal trend. As we have seen, detrending reduces the autocorrelation and therefore the probability of false alarm.

[2]The numbers given are mean $\pm$ one standard deviation.

[3]Obtained as the difference in Mon-Fri and Sat-Sun means, 7.81 ppm $-(-0.81)$ ppm.

[4]These are given as mean $\pm$ 2$\times$(standard deviation).

Mauna Loa observatory's remote location.

The above discussion suggests the a shift is not detectable at Mauna Loa, and that a bigger shift is detectable at the urban University of Utah site than the rural Harvard Forest site. The signal to noise ratios in Table 7.1 support this trend. The signal to noise ratios of the weekend weekday shift are greatest (on winter mornings) for Utah (0.278), smaller for Harvard Forest (0.144), and near zero for Mauna Loa (0.027).

It is interesting to think about what would happen if all times were included. The relatively smaller signal to noise ratios obtained when considering all times suggest including all residuals in the analysis means diluting the shift we are looking for, hampering the ability to detect the shift by including residuals that are not sensitive to shifts in urban emissions.

7.1.3 The need for a Bayesian detection method

Comparisons based on 95% confidence intervals assume the residuals are normally distributed. Figure 7-4 compares the weekend and weekday distributions for the residuals from each site against normal distributions. It is clear the CO_2 residuals (and hence forecasts) are non-Gaussian. This conclusion is confirmed by normality tests. The chi-square, Lilliefors, and Jarque-Bera tests outlined in Section 2.8.3 all reject normality of the Winter AM residuals from the sites considered at the 0.1% level of significance. Standard detection methods provide poor control of false positive rates when forecast uncertainties are non-Gaussian. The next section therefore introduces a Bayesian approach to detection that can handle non-Gaussian distributions.

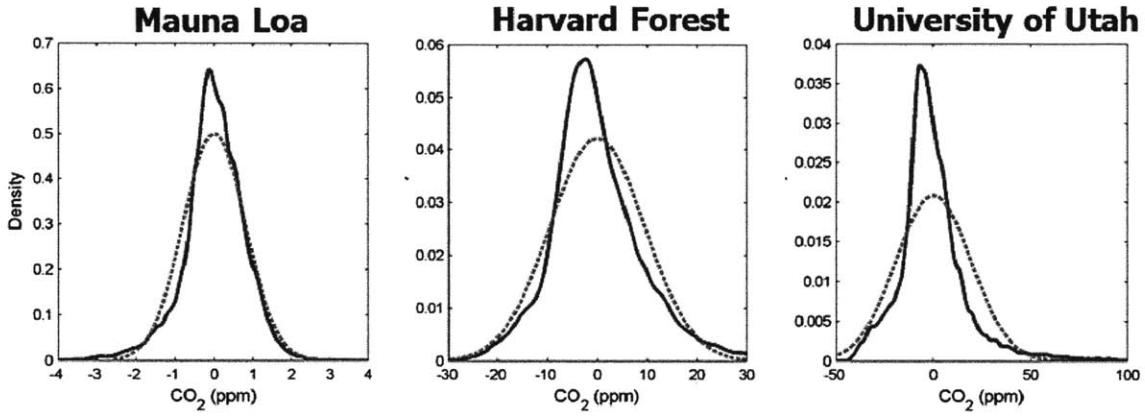

Figure 7-4: Comparison of residual distributions from Model SRO (blue) with Gaussian fits (dashed red).

7.2 Shift detection using Bayes factors

The previous section demonstrated the need for Bayesian methods capable of handling non-Gaussian uncertainties. The concept behind the method introduced in this section was introduced earlier in Figure 2-5. The idea is to detect a shift in the distribution of the data. Figure 7-5 illustrates how this idea applies to detecting weekend weekday shifts in atmospheric CO_2.

7.2.1 Bayes' theorem and Bayes factors

The challenge is assessing which distribution to believe in given new data. The discrete version of Bayes' theorem provides a statistically sound basis for addressing this problem. The theorem was introduced earlier in (2.39). It is repeated here with

Table 7.1: Weekend weekday CO_2 shift statistics

Mauna Loa

Subset		n_{obs}	Mean (ppm)	Stdev of residuals (ppm)	Stdev of mean (ppm)	Shift SNR[c]
All	Sat-Sun	15,912	-0.110	0.906	0.007	0.054
	Mon-Fri	33,392	-0.062	0.890	0.005	
Summer	Sat-Sun	8,021	-0.090	1.020	0.011	0.061
	Mon-Fri	16,764	-0.029	0.989	0.008	
Winter	Sat-Sun	7,891	-0.130	0.773	0.009	0.046
	Mon-Fri	16,628	-0.094	0.777	0.006	
Summer AM	Sat-Sun	1,679	0.089	0.889	0.022	0.094
	Mon-Fri	3,749	0.171	0.856	0.014	
Winter AM	Sat-Sun	1,647	-0.027	0.725	0.018	0.027
	Mon-Fri	3,699	-0.008	0.701	0.012	
Summer PM	Sat-Sun	1,672	0.310	0.878	0.021	-0.005
	Mon-Fri	3,652	0.305	0.888	0.015	
Winter PM	Sat-Sun	1,648	0.035	0.710	0.017	0.060
	Mon-Fri	3,631	0.077	0.703	0.012	

Harvard Forest

Subset		n_{obs}	Mean (ppm)	Stdev of residuals (ppm)	Stdev of mean (ppm)	Shift SNR[c]
All	Sat-Sun	21,466	-2.369	9.400	0.064	0.048
	Mon-Fri	50,626	-1.905	9.847	0.044	
Summer	Sat-Sun	12,088	-2.775	10.639	0.097	-0.008
	Mon-Fri	28,671	-2.860	10.864	0.064	
Winter	Sat-Sun	9,378	-1.847	7.476	0.077	0.152
	Mon-Fri	21,955	-0.658	8.169	0.055	
Summer AM	Sat-Sun	3,005	-3.379	9.727	0.177	-0.023
	Mon-Fri	6,929	-3.599	9.538	0.115	
Winter AM	Sat-Sun	2,275	-1.588	7.637	0.160	0.144
	Mon-Fri	5,339	-0.459	8.046	0.110	
Summer PM	Sat-Sun	3,035	-10.848	7.801	0.142	-0.054
	Mon-Fri	7,300	-11.269	7.859	0.092	
Winter PM	Sat-Sun	2,369	-3.656	7.073	0.145	0.138
	Mon-Fri	5,415	-2.626	7.809	0.106	

Utah

Subset		n_{obs}	Mean (ppm)	Stdev of residuals (ppm)	Stdev of mean (ppm) [b]	Shift SNR[c]
All	Sat-Sun	15,758	-4.97	18.07	0.144	0.158
	Mon-Fri	34,609	-2.07	18.73	0.101	
Summer	Sat-Sun	7,718	-3.02	8.34	0.095	0.188
	Mon-Fri	17,100	-1.34	9.52	0.073	
Winter	Sat-Sun	8,040	-6.85	23.80	0.265	0.168
	Mon-Fri	17,509	-2.78	24.57	0.186	
Summer AM	Sat-Sun	1,614	0.36	8.84	0.220	0.529
	Mon-Fri	3,537	5.89	11.83	0.199	
Winter AM	Sat-Sun	1,674	-0.81	27.69	0.677	0.278
	Mon-Fri	3,649	7.05	28.95	0.479	
Summer PM	Sat-Sun	1,598	-9.56	4.68	0.117	0.290
	Mon-Fri	3,557	-8.17	4.91	0.082	
Winter PM	Sat-Sun	1,675	-14.00	18.82	0.460	0.187
	Mon-Fri	3,642	-10.39	19.81	0.328	

[a] Winter = Oct-Mar, Summer=Apr-Sep, AM = 6-10am, PM = 12-4pm.
[b] Calculated in the usual way, as Stdev of residuals divided by square root of n_{obs}.
[c] Defined as (wkday mean - wkend mean)/(0.5*var of res for wkend + 0.5*var of res for wkday). See (3.3).

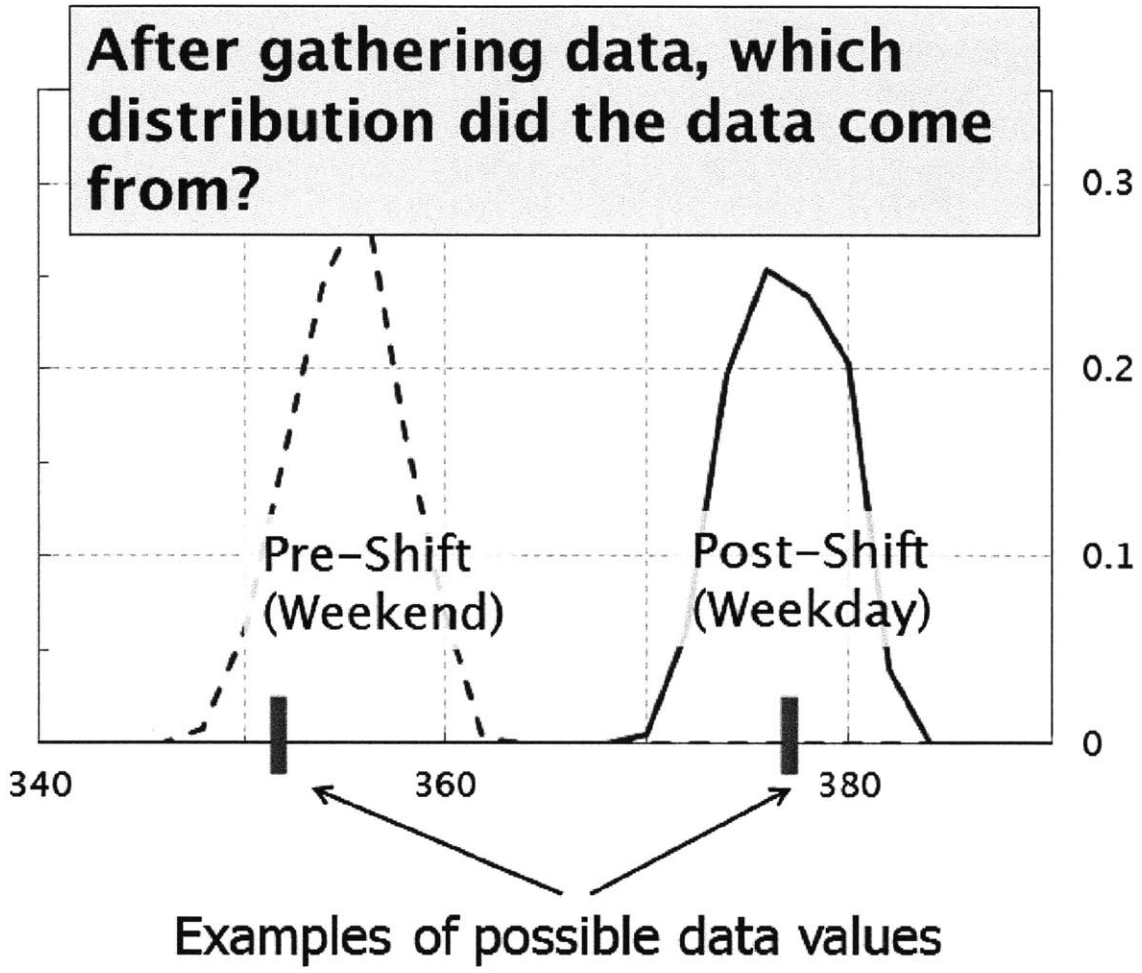

Figure 7-5: Concept for Bayesian detection of weekend weekday shifts in CO_2. The curves shown represent the distributions of the data before and after the shift (that is, on weekends and weekdays). The challenge is assessing which distribution to believe in given new data.

slightly different notation appropriate for the shift detection problem:

$$P(H_i|y) = \frac{f(y|H_i)P(H_i)}{f(y)} \quad i = \begin{cases} 0 & \text{if } y \text{ comes from Saturday or Sunday, and} \\ 1 & \text{otherwise.} \end{cases} \tag{7.1}$$

$P(H_i)$ is the prior probability the new data being observed comes from a weekend or weekday. Bayes' theorem updates the prior probability after new data (y) is observed. The new probability is the posterior $P(H_i|y)$. The bigger the probability associated with a model, the more likely it is to be the target model. $f(y)$, the density of the data marginalized over the models, is independent of the hypothesis of interest and thus can be ignored. The data enter through the likelihood $f(y|H_i)$, which is the predictive density function of the data associated with hypothesis H_i. The cartoon density functions in Figure 7-5 illustrate the likelihoods for the weekend and weekday hypotheses. Given an observation, the values implied for each of the competing densities are what enter Bayes' theorem.

The posterior odds in favor of H_0 is defined as the ratio of posterior probabilities

$$\frac{P(H_0|\boldsymbol{y})}{P(H_1|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|H_0)P(H_0)}{f(\boldsymbol{y}|H_1)P(H_1)} \tag{7.2}$$

The ratio of the predictive densities is called the Bayes factor

$$BF = \frac{f(\boldsymbol{y}|H_0)}{f(\boldsymbol{y}|H_1)} \tag{7.3}$$

which is just a ratio of predictive densities. When the researcher prefers neither hypothesis over the other a priori, then the posterior odds ratio is equivalent to the Bayes factor:

$$\frac{P(H_0|\boldsymbol{y})}{P(H_1|\boldsymbol{y})} = BF \tag{7.4}$$

Thus, the Bayes factor conveys information about which hypothesis is preferred, given the new data $\boldsymbol{y}$. Combined with prior odds ($P(H_0)/P(H_1)$), the Bayes factor enables the evaluation of the posterior odds in favor of H_0.

Evaluating the Bayes factor requires evaluating predictive densities (see Appendix A). This requires integrating over the distributions of the parameters and hyperparameters affecting forecasted values of $\boldsymbol{y}$. For sake of example, suppose the model parameters, observation errors, and misfit errors are all Gaussian and independent of time, so

$$\begin{aligned} &\boldsymbol{\theta} \sim N(\hat{\boldsymbol{\theta}}, \boldsymbol{V}_\theta), \\ &\varepsilon_{obs}(t) \sim N(0, \sigma^2_{obs}), \text{ and} \\ &\varepsilon_{\mathcal{M}}(t) \sim N(0, \sigma^2_{\mathcal{M}}) \end{aligned} \tag{7.5}$$

where σ^2_{obs} and $\sigma^2_{\mathcal{M}}$ are themselves uncertain according to

$$\begin{aligned} \sigma^2_{obs} &\sim N(a_0, a_1^2) \text{ and} \\ \sigma^2_{\mathcal{M}} &\sim N(b_0, b_1^2). \end{aligned} \tag{7.6}$$

The a's and b's are hypothesis-dependent hyperparameters dictating the scale of the observational and misfit errors. In this case evaluating the predictive density for H_i requires evaluation of the integral

$$BF = \frac{\int f(\boldsymbol{y}|\boldsymbol{\theta}, a_0, a_1, b_0, b_1, H_0)\pi(\boldsymbol{\theta}, a_0, a_1, b_0, b_1|H_0)d\boldsymbol{\theta}da_0da_1db_0db_1}{\int f(\boldsymbol{y}|\boldsymbol{\theta}, a_0, a_1, b_0, b_1, H_1)\pi(\boldsymbol{\theta}, a_0, a_1, b_0, b_1|H_1)d\boldsymbol{\theta}da_0da_1db_0db_1} \tag{7.7}$$

Forward and observational uncertainties were demonstrateed earlier to be negligible compared to misfit uncertainty. Therefore $\boldsymbol{\theta}$ can be treated as if it were known exactly (with value $\hat{\boldsymbol{\theta}}$ obtained from ordinary least squares) known and it is safe to assume $a_0 = 0$ and $a_1 = 0$ so that observational uncertainty drops out. The only parameters we need to worry about are b_0 and b_1, which characterize the uncertainty in the residuals. Since thousands of residuals are used to formulate the weekend and weekday distributions in this chapter, these parameters might be treated as if they were exact. These assumptions simplify (7.7) to

$$BF = \frac{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}, H_0)}{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}, H_1)} \tag{7.8}$$

Thus, the Bayes factor simplifies to a traditional likelihood ratio.

This motivates the following conceptual procedure.

1. The detrending parameters are estimated by ordinary least squares, resulting in a site specific value $\hat{\boldsymbol{\theta}}$.

2. This value is used to detrend observations, resulting in residuals $\boldsymbol{e}$.

3. Two subsets of the residuals are created, one including weekend values and the other weekday values.

   - The weekend subset defines the population of weekend model errors, with associated predictive density $f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}, H_0)$.
   - The weekday subset defines the population of weekday model errors, with associated predictive density $f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}, H_1)$.

4. The posterior odds after observing new data, $\boldsymbol{y}_{obs}$, is updated by the Bayes factor obtained by inserting the new data into (7.8).

Two types of predictive distributions are considered:

- Empirical density. The empirical density function is close to the actual distribution of the residuals. The empirical densities used here are based on kernel density estimators.

- Normal density. These are Gaussian approximations to the true distribution, which based on the normality test results described earlier are different from the actual distribution at the 0.001 level of significance.

The next section gives a more detailed algorithm for analysis based on Bayes factors computed as explained above.

7.2.2 Bayesian detection algorithm

The procedure is as follows:

1. Distributions. Because the algorithm is Bayesian, the densities $f(\boldsymbol{y}|\theta_i', H_i)$ could take any form supported by the residuals. We considered two possible forms. The first was a normal fit to the residuals. The second was a Gaussian kernel density estimator. Sample means and variances are sufficient to characterize normal fits. Kernel density estimators can be obtained in MATLAB using the *ksdensity* function.

2. Resampling. Two ensembles of random trajectories are generated by resampling the null and alternative densities $f(Y|\theta_0')$ and $f(Y|\theta_1')$. The null ensemble is a set of $n_{obs} \times 1$ vectors $\{\boldsymbol{y}_0^{(1)}, \boldsymbol{y}_0^{(2)}, ..., \boldsymbol{y}_0^{(n_s im)}\}$. The alternative ensemble was a similar set of $n_{obs} \times 1$ vectors $\{\boldsymbol{y}_0^{(1)}, \boldsymbol{y}_0^{(2)}, ..., \boldsymbol{y}_0^{(n_s im)}\}$. To simplify our notation, we denote each of these ensembles as $\boldsymbol{Y}_i = \{y_{ij}^{(k)}\}$, where i denotes the ensemble for hypothesis H_i $(i = 0, 1)$, j denotes the j^{th} value in a given trajectory $(j = 1, 2, ..., n_{obs})$, and k denotes the k^{th} trajectory in the ensemble $(k = 1, 2, ..., n_{sim})$. $n_s im$ was set to 10,000. The number of observations can be varied to simulate different sized detection windows.

3. Bayes factors. The samples are assumed independent, so the Bayes factor for trajectory k generated under hypothesis H_i can be computed according to

$$BF_i^{(k)} = \prod_{j=1}^{nobs} \frac{f(y_{ij}^{(k)}|\theta_0', H_0)}{f(y_{ij}^{(k)}|\theta_1', H_1)} \tag{7.9}$$

   To avoid numerical problems, the log Bayes factor was first generated and then exponentiated to arrive at the desired Bayes factor. Therefore,

$$BF_i^{(k)} = \exp\left[\sum_{j=1}^{nobs} \ln f(y_{ij}^{(k)}|\theta_0', H_0) - \log f(y_{ij}^{(k)}|\theta_1', H_1)\right] \tag{7.10}$$

   The result was two ensembles of Bayes factors. The null ensemble is $BF_0^{(1)}, ..., BF_0^{(n_s im)}$. The alternative ensemble is $BF_1^{(1)}, ..., BF_1^{(n_s im)}$. We represent these ensembles as $BF(\boldsymbol{Y}_i)$ to highlight the dependence on the observational ensemble $\boldsymbol{Y}_i$.

4. Probabilities. We specified the prior weights for H_0 and H_1 as $w_0 = w_1 = 0.5$. Thus, the decision criterion is a maximum a posteriori (MAP) criterion, which

chooses H_0 if the value of the Bayes factor is greater than 1, and chooses H_1 otherwise. Probabilities are then obtained as follows:

Detection probability

$$P_d = \text{Prob}\left(\text{BF}\left(\mathbf{y}\right) \leq 1 | \text{H}_1\right) = \frac{\text{No. of BF's in } BF(Y_1) \text{ less than or equal to } 1}{n_{sim}} \tag{7.11}$$

False alarm probability

$$P_{fa} = \text{Prob}\left(\text{BF}\left(\mathbf{y}\right) \leq 1 | \text{H}_0\right) = \frac{\text{No. of BF's in } BF(Y_0) \text{ less than or equal to } 1}{n_{sim}} \tag{7.12}$$

The procedure outlined above is illustrated in Figure 7-6. A key idea is the fact that as the number of samples available increases, the null and alternative Bayes factor distributions become more distinct. This fact, illustrated in Figure 7-7, is what enables us to attain the same detection rate at reduced frequencies of false alarm.

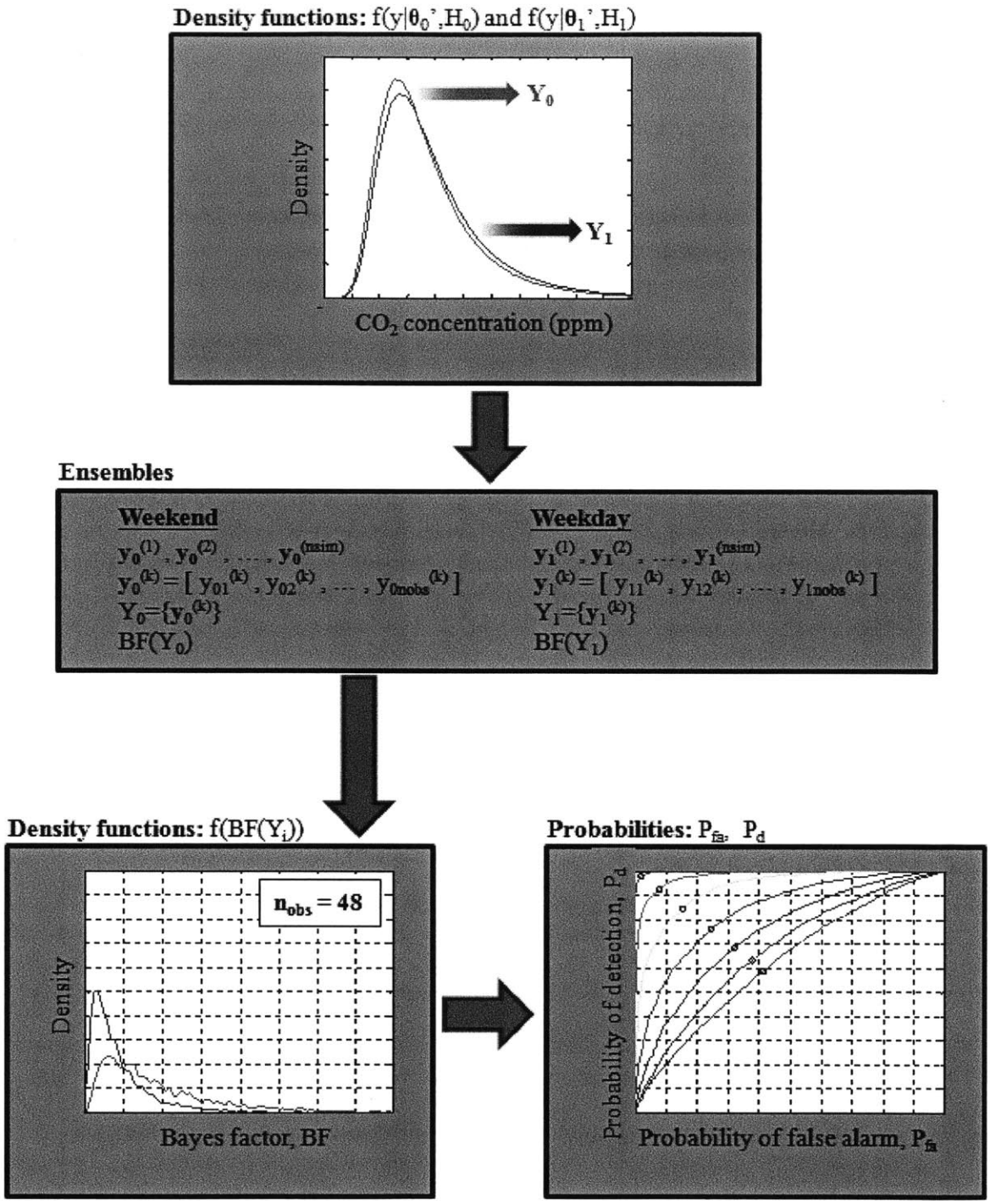

Figure 7-6: Illustration of the Bayesian detection algorithm.

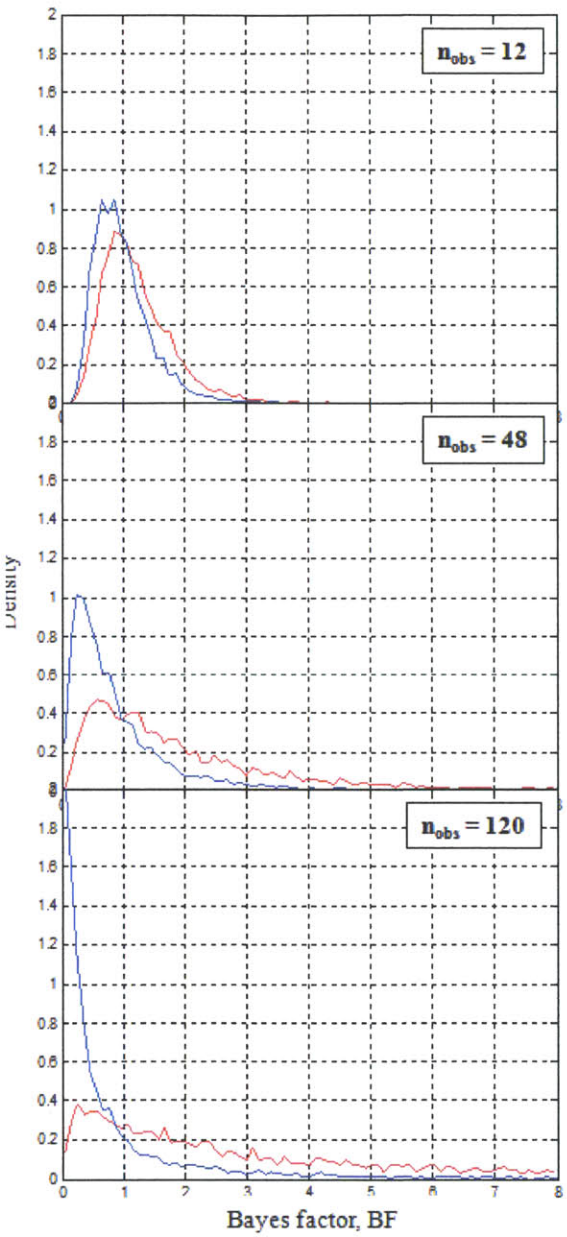

Figure 7-7: Effect of growing sample size on Bayes factor densities. The blue curves depict the Bayes factor density given samples from the weekday population. The red curves are the same given samples come from the weekend population.

7.2.3 Results for Harvard Forest and Salt Lake

We used the algorithm described in Sections 7.2 and 7.2.2 to determine how many samples (N) would be required to detect the weekend weekday shifts at Harvard Forest and Utah at the 0.05 level of significance. This section describes the findings. The analysis was restricted to winter months (Oct-Mar) between 6 a.m. and 10 a.m. in the morning. The detrending model used was SRO, which has the form

$$c(t) = \theta_1 + \theta_2(t-t_0) + \theta_3 SOLRAD(t-\theta_4) + \theta_9 \cos[4\pi(t-t_0)/365] + \theta_{10} \sin[4\pi(t-t_0)/365] \tag{7.13}$$

The parameter estimates $\hat{\boldsymbol{\theta}}$ were obtained using ordinary least squares, considering the Training data sets defined in Table 5.3. The parameter estimates for Harvard Forest and University of Utah are provided in Table 7.2.

Table 7.2: Least squares parameter estimates for model SRO

Site	θ_1	θ_2	θ_3	θ_4	θ_9	θ_{10}
Harvard	379.5±0.4	0.0070±0.0002	0.018±0.000	33.3±0.8	-3.7±0.1	-1.2±0.1
Utah	439.3±0.5	0.0080±0.0004	0.048±0.000	23.3±0.6	0.3±0.2	6.4±0.2

Residuals were obtained by considering six years of hourly mean atmospheric CO_2 data from Salt Lake's University site and twelve years[5] from Harvard Forest. At Harvard Forest, there were 2,275 weekend and 5,339 weekday samples available. At Utah there were 1,674 and 3,649 of the same, respectively.

Detecion results were obtained using both normal and empirical densities to test the robustness of the algorithm to the normality assumption

The following figure depicts the weekend and weekday predictive densities (for residuals) at Harvard Forest and University. The densities conform with what we expet in that they appear to be non-Gaussian. There are slight upward shifts in the residuals from weekends to weekdays at both sites, with the shift at Utah being more obvious. These are presumably due to changes in urban emissions.

[5]More years were used at Harvard Forest to improve the ability to detect the smaller shift expected there.

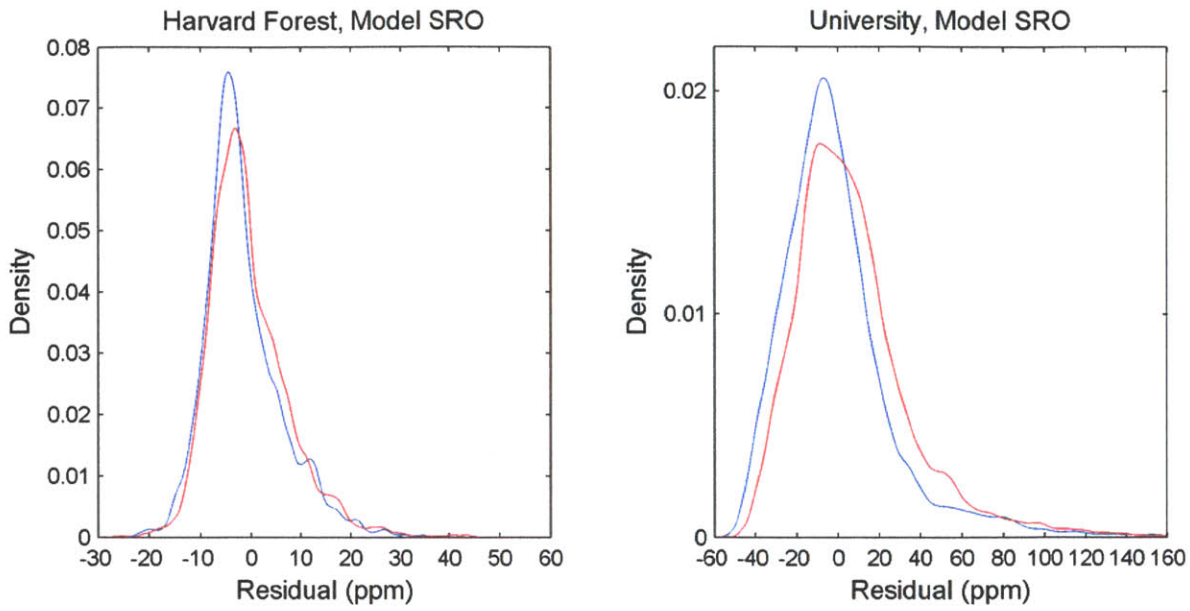

Figure 7-8: The weekend weekday shift at Harvard Forest and University of Utah. The curves shown are kernel density fits to weekend (blue) and weekday (red) residuals obtained by detrending the Training data set for each site using the least squares estimates for model SRO.

Earlier analysis based on Table 7.1 showed that the mean shift size is roughly +1.13 ppm at Harvard Forest and +7.86 ppm at the University of Utah. The shifts are positive, supporting the hypothesis that increased activity during the week is leading to increased fossil fuel emissions. The signal to noise ratio of the weekend weekday shift is bigger at Utah than Harvard Forest, in accord with expectations that the signal at the semi-rural Harvard Forest site be weaker than at the urban University site. This is despite much larger diurnal deviations at Utah due to urban emissions.

The weekend weekday urban CO_2 shift detection results are summarized in Tables 7.3 and Figure 7-9. Time to detection was computed as the number of samples required to reduce the false alarm rate[6] to below 5%.

The results suggest time to detection is sensitive to the distribution assumed for the residuals (empirical versus normal). Thus, results based on empirical distributions should be considered. The shift at Harvard is detectable at the 0.05 level of significance in 120-240 samples. The shift at Utah is detectable at the 0.05 level of significance in 48-120 samples, fewer samples as expected given the bigger signal to noise ratio at that site. If samples are recorded hourly, at least 2 days at Utah and 5 days of data at Harvard Forest should be considered to detect at the 0.05 level of significance.

Table 7.3: Time to detection at University

No. samples	Normal		Empirical	
	P_{fa}	P_d	P_{fa}	P_d
12	0.3100	0.6780	0.2740	0.7440
24	0.2440	0.7460	0.2010	0.8330
48	0.1590	0.8280	0.1050	0.9100
120	0.0610	0.9250	0.0200	0.9890
240	0.0120	0.9800	0.0010	0.9970
480	0.0000	0.9990	0.0000	1.0000

[6]The false alarm rate associated with MAP-based (maximum a posteriori based) detection.

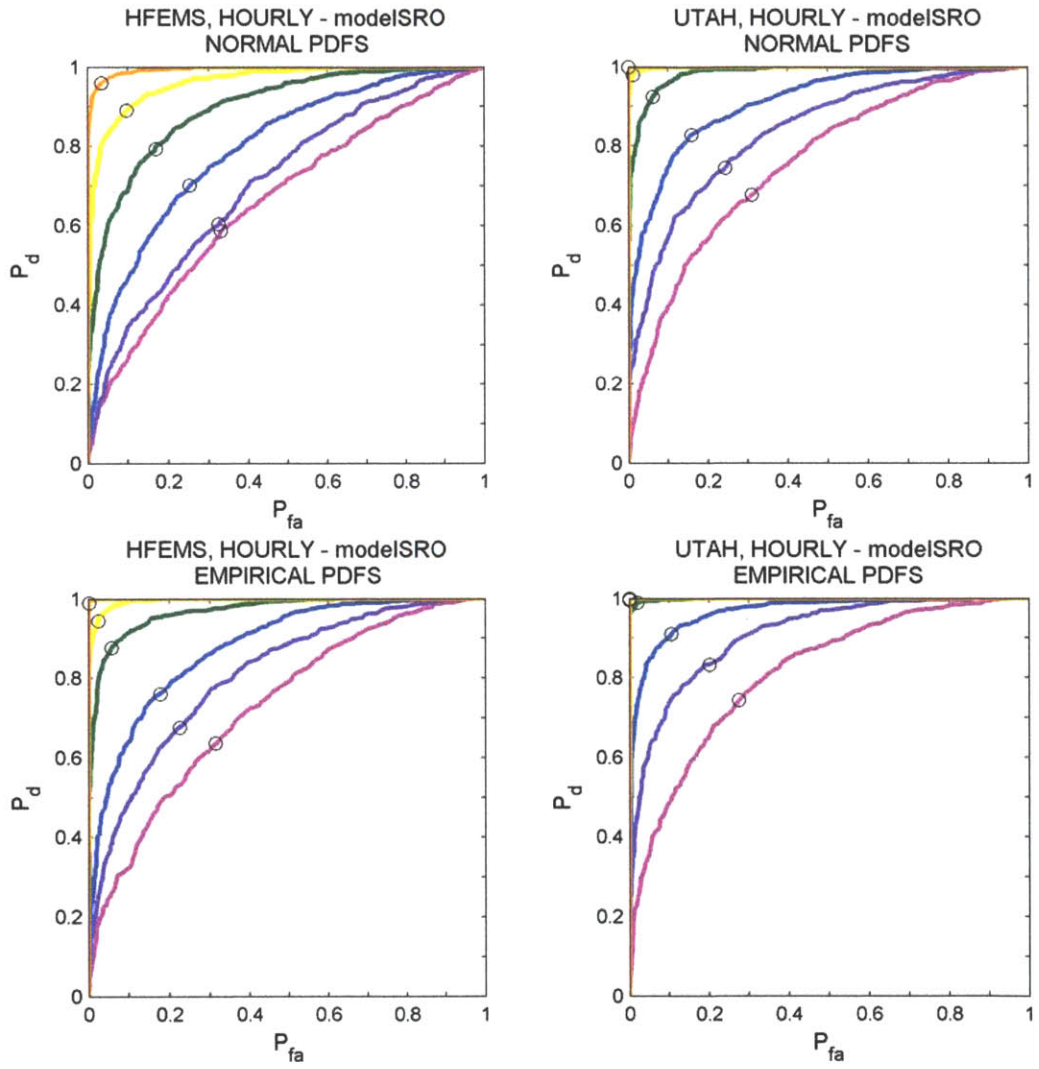

Figure 7-9: Probabilities of detection and false alarm for the weekend weekday shift at Harvard Forest and University of Utah. In each figure, the curves shown are the results from 10^3 simulations of trajectories with 12, 24, 48, 120, and 480 samples from either normal (left) or empirical (right) densities. The kernel densities alluded to are the ones shown in Figure 7-8.

Table 7.4: Time to detection at Harvard Forest

	Normal		Empirical	
No. samples	P_{fa}	P_d	P_{fa}	P_d
12	0.330	0.588	0.314	0.637
24	0.324	0.605	0.225	0.677
48	0.252	0.702	0.177	0.761
120	0.168	0.794	0.057	0.877
240	0.096	0.89	0.024	0.944
480	0.035	0.96	0.002	0.989

7.3 Lessons from Phoenix, AZ

The urban effect on atmospheric CO_2 has been considered previously. One set of studies began in the late 1990's in Phoenix, Arizona [89, 91, 92]. The intensive measurements conducted in the Phoenix studies established the presence of an urban CO_2 "dome", a plume of greatly enhanced atmospheric CO_2, generated by fossil fuel combustion in the city.

7.3.1 Variability in the Phoenix CO_2 dome

Idso et al. [89] improved on an earlier study by the same authors by making additional measurements of atmospheric CO_2 along four transects in and around Phoenix. Measurements were made at two meters above the surface and at one mile intervals. Two transects passed directly through the city center, and two traveled the city perimeter. Each day for two weeks, runs were initiated at 5am and 2pm. The data enable several observations: (1) atmospheric CO_2 varies greatly in residential areas, with differences of $\sim$ O(70ppm) present in the transects traversing the city perimeter. (2) Greater variability is observed on the transects cutting through the city center, with concentrations $\sim$ O(140ppm) higher in the city center than the surroundings. (3) Morning-afternoon differences as large as $\sim$ O(70ppm) were observed with morning concentrations typically higher, highlighting the importance of diurnal variability and vertical mixing. (4) The authors note increased mean concentrations on weekdays compared to weekends near the city center.

Idso et al. [91] followed by making using high frequency (one minute average), continuous in situ CO_2 measurements from two meters height in a residential backyard roughly 20 km southeast of the city center. The high frequency, continuous data enabled the authors to focus on the times of day when the atmospheric CO_2 signal was most sensitive to urban emissions. The diurnal trend in winter was observed to differ from that in the summer, with both early morning and late evening peaks in the winter but only early morning peaks in the summer. The authors estimated that the urban signal in the city center was $\sim$ O(250ppm) and still $\sim$ O(120ppm) as far as

20 km away. As in their previous study, Idso et al. noted increased atmospheric CO_2 on weekdays over weekends, but in addition they were able to highlight the times of day when the shift would be easiest to detect. In the summer, they conclude that the weekend weekday shift is at most 22.0 ppm at 7:09am, with the mean difference being positive between 4:15 and 8:30am. In the winter, they conclude that the weekend weekday shift is at most 35.9 ppm at 7:47am, with the mean difference being positive between 4:45 and 10:45am. Although the authors did not discuss it, the later peak shift time in the winter is interesting because of the possible role of daylight savings.

Koerner and Klopatek [92] generated bottom-up estimates of CO_2 emissions in Phoenix and found that roughly 80% of emissions were attributable to vehicles and the remainder were primarily due to soil respiration. In total, they estimate total annual emissions were roughly $\sim$ 93Mg/yr in 1997.

7.3.2 Estimating the weekend weekday emissions shift in Phoenix, Arizona

According to the 1990 United States Census Bureau metropolitan area predictions used by Koerner and Klopatek [93], the population of Phoenix, AZ, was forecasted to be 3.013 million in 1999 while that of Salt Lake City was forecasted to be just 1.275 million. Phoenix's population in 1990 was approximately 2.238 million. Interpolation of the Census Bureau data predicts Phoenix's population to have been roughly $2.238 + 7/9 \cdot (3.013 - 2.238) \approx 2.8$ million people in 1997. Using the vehicle emissions estimates obtained by Koerner and Klopatek, this implies per capita vehicle emissions of 67.7 $MtCO_2yr^{-1}/2.8$ million people $\approx$ 24 tonnes CO_2yr^{-1} per person.

To check that this makes sense, assume the typical gas tank holds about 100 kg of gasoline (25 gallons x 4 kg per gallon) and that gasoline is three times lighter than CO_2 (the latter based on the ratio of molecular weights of CO_2 to CH_2.) Then 24,000 kg CO_2 works out to 8,000 kg of gasoline or 80 trips to the gas pump. This means each person is refilling their gas tank every 4-5 days, which seems reasonable.

Next, consider the shifts detected by Idso et al. [91]. They detect shifts on the order of 20-35 ppm at a site 20 km away from an estimated urban peak signal of 250 ppm.

Judging by the map provided by Koerner and Klopatek, the Phoenix metropolitan area was $\sim$800 km^2 at the time of the study. This means a mean vehicular flux of $6{,}770 \times 10^4 tCO_2yr^{-1}/800km^2$ or roughly 80 $ktCO_2km^{-2}yr^{-1}$. This is the average for the entire year. However, assuming a linear relationship between atmospheric concentration and vehicular flux, taking 80% of 20-35 ppm (= 16 to 28 ppm) to represent the weekend weekday shift in atmospheric CO_2 due to traffic, and taking 80% of 250 ppm (= 200 ppm) to be the peak urban plume concentration in the absence of sources other than traffic during the week when traffic is greatest, we can solve for the weekend and weekday fluxes of CO_2 by solving the following system of

equations:

$$\begin{aligned} \frac{2}{7}F_{wkend} + \frac{5}{7}F_{wkday} &= 80 \text{ ktCO}_2\text{km}^{-2}\text{yr}^{-1} \\ \frac{F_{wkday}}{F_{wkend}} &= \frac{200 \text{ ppm}}{200 - \Delta y \text{ ppm}} \end{aligned} \tag{7.14}$$

where Δy denotes the magnitude of the weekend weekday shift in CO_2 concentration in ppm. Thus, using the summer-derived shift of 16 ppm, we obtain $F_{wkend} = 75.3$ and $F_{wkday} = 81.9$, so the 16 ppm shift translates into a flux change of +6.6 $\text{ktCO}_2\text{km}^{-2}\text{yr}^{-1}$. Using the larger shift value of 28 ppm associated with the winter yields $F_{wkend} = 71.7$ and $F_{wkday} = 83.3$ for a flux change of +11.6 $\text{ktCO}_2\text{km}^{-2}\text{yr}^{-1}$. Thus, the weekend weekday "flux" (resulting from increased emissions during the week) in Phoenix is ~10 $\text{ktCO}_2\text{km}^{-2}\text{yr}^{-1}$.

7.3.3 Extrapolating to Salt Lake City, Utah

What size shift can we expect to observe in Salt Lake City, UT? Let us assume that per capita vehicular emissions are the same in Salt Lake and Phoenix, which based on the numbers already presented is roughly 24,000 $\text{kgCO}_2\text{yr}^{-1}$ per person. According to the U.S. Census Bureau's latest statistics on metropolitan and micropolitan areas [94], the population of Salt Lake City is estimated to have grown from about 0.97 million in 2000 to 1.13 million people in 2009. Furthermore, Salt Lake City is contained in a valley that is roughly 10km wide × 30km long, giving a total land area of about 300 km^2. Vehicular fluxes can be obtained by taking the population value, multiplying the result by 24,000 $\text{kgCO}_2\text{yr}^{-1}$ per person, and dividing by 300 km^2. This predicts an annual vehicular flux in Salt Lake that grows as the population increases from 77.6 to 90.4 $\text{kgCO}_2\text{km}^{-2}\text{yr}^{-1}$ between 2000 and 2009, respectively. Even though the population of Salt Lake City is much smaller than that of Phoenix, the vehicular fluxes are of similar magnitude because Salt Lake City's population is contained in a much smaller area.

We expect the magnitude of the weekend-weekday shift in CO_2 emissions to be comparable to that in Phoenix, ~ 10 $\text{ktCO}_2\text{km}^{-2}\text{yr}^{-1}$. Barring any major differences in transport, this should translate into similarly large enhancements in ambient CO_2 levels. It is therefore expected that (1) the CO_2 plume in Salt Lake City should have a peak of the similar order of magnitude ($O(300)$ ppm) to the one in Phoenix [91] and (2) that weekend weekday shifts in atmospheric CO_2 should also be of similar magnitude ($O(30)$ ppm).

At first glance, the first of the expectations for atmospheric CO_2 appears to be met - concentrations between 500 and 700 ppm are routinely observed in urban Salt Lake city. Concentration at the rural site tend to be less extreme, only reaching up to ~450 ppm. The second hypothesis is less obvious. Figure 7-3 gives 24-hour plots of hourly mean atmospheric CO_2 at the University site in Salt Lake City. The figure suggests the presence of upward shifts in CO_2 in some, but not all, months and that the shift should be more significant in the morning when atmospheric mixing is weak

but emissions from rush hour traffic are substantial.

7.4 Implications for leak detection for geological sequestration

The earlier comparison between Phoenix and Salt Lake City suggests the urban weekend weekday shift detected at Utah is the result of a "flux" of $\sim O(0.01)$ $MtCO_2km^{-2}yr^{-1}$. The biggest scale sequestration projects today like Sleipner inject $\sim O(1)$ $MtCO_2yr^{-1}$ belowground. Thus, it should be possible to detect a 1% leak rate[7] occurring over 1 km^2, simply by removing the seasonal signal from atmospheric CO_2 concentrations downwind of the source and testing the residuals for an upward mean shift. Detecting the shift will become difficult the farther the sensor is from the leak, as illustrated through the Harvard Forest example. Detecting leakage of smaller magnitude or over larger areas will decrease the signal to noise ratio, and may require improving models to explain higher frequency, diurnal effects.

[7]Defined as 1% of the amount stored per year.

Chapter 8

Procedure for detecting shifts in CO_2

The issues and analysis presented in the previous chapters suggest the following procedure for monitoring at new sites. First, baseline data are collected and models for the data proposed. The data are split into two subsets. A training subset is used to estimate model parameters. The estimates are then used on the validation data set with a model selection procedure such as DIC to select the best model. The posterior uncertainty in the detrending parameters and the scale of variability in the residuals determine uncertainty in future CO_2 data. The weekend weekday shift detection approach can be applied to artificial shifts in the data to determine how much data needs to be collected to detect shifts associated with future leaks at the desired level of significance. If a single sensor is being used to monitor, then the number of data samples (N) required to detect and the sampling interval ($N\Delta t$) dictate the size of the window to use to analyze for shifts.

Figure 8-1 gives a flow chart for the procedure. The researcher sets a false alarm rate (α) and chooses a model ($\mathcal{M}$). Simulations of hypothetical leaks from the reservoir can be used to provide worst-case scenario estimates of CO_2 shifts Δy; due to uncertainty in subsurface and atmospheric transport, the shift arriving at the sensor will be uncertain and can be characterized by a density function $f(\Delta y)$. The model's ability to predict past data determines the scale of the model errors ($\varepsilon_{\mathcal{M}} + \varepsilon_{obs}$); uncertainty in the parameters adds uncertainty to the forecast through the detrending model (ε_{theta}). The forecast uncertainty, uncertainty about the size shift arriving in the future ($f(\Delta y)$) and the false alarm rate selected by the researcher will determine the window size (T) used to analyze the data for shifts. Once a window size has been chosen, new observations for the next T days are gathered. These are tested for a shift. If no shift exists, then the observations are added to the baseline and the current time is stepped forward, awaiting the next T days' observations. Otherwise a shift exists, and steps should be taken to perform remediation. As time progresses, the processes affecting the observations may change.

Figure 8-1: Detecting shifts in CO_2 in practice.

Chapter 9

Conclusions

9.1 The leak detection problem

Carbon capture and sequestration has been proposed as a way to remove CO_2 from the atmosphere by injecting it into deep saline aquifers. It is important to ensure that the CO_2 remains underground and does not leak into the atmosphere. Monitoring and verification are needed for detection and mitigation of potential leaks. A standard set of tools for monitoring and verifying leaks to the atmosphere does not yet exist. Important questions remain unanswered. For example, what is the smallest leak detectable? And how much time is needed to detect a leak?

The ultimate goal for monitoring and verification at geological sequestration sites is determining when a leak occurs and how its magnitude evolves over time. Many measurement techniques exist for inferring sources and sinks of CO_2. Eddy covariance is useful when the leak occurs within a small footprint of the sensor. Atmospheric dispersion methods are useful for inferring fluxes farther upwind based on perturbations in downwind concentrations. The basic problem is estimating these perturbations.

The most basic form of the leak detection problem is to detect and estimate upward shifts in CO_2 concentration time series like the Keeling curve. The solution is complicated because CO_2 varies widely in time and space. Predictive models can help by enabling forecasts conditioned on knowledge of past trends. When a new set of observations differs from such a forecast, a leak may be suspected. Mean shift detection methods can be used to determine if a shift has occurred. If a shift has occurred, then the population can be alerted and the source mitigated.

The leak detection problem is a mean shift detection problem. Mean shifts arise under many different names in the literature [23], including observational inhomogeneities, change points, and regime shifts. Homogenization is the practice of detecting and then correcting for artificial shifts in time series and is commonly used in climate science. Regime shift detection methods arise in finance, economics and ecology where there is interest in issues such as the onset of economic recession and changes in population dynamics. The changes being sought out can be mean shifts, shifts in variance, changes in harmonic content, or structural changes in the system governing the observations [24]. Statistical approaches to detection are useful because

forecasts are uncertain. Imperfect instrumental calibration, averaging, and complexity of the physical system imply predictions of future CO_2 will always be prone to error. This error implies uncertainty about future observations. The challenge is therefore detecting a shift from an uncertain quantity.

Not all statistical mean shift detection methods are applicable to the leak detection problem. Some of the issues are:

- The time when a leak arises is unknown, so the time at which CO_2 shifts upward (t_s) is unknown.
- The magnitude of the CO_2 shift we are looking for (Δy) is unknown.
- The model errors may not be Gaussian.
- The model errors may be correlated.
- Gaps are typically present in the data.

Some of the most familiar detection methods, including heuristics like RSS, significance tests like Student's t-test, and likelihood based tests, are poorly suited for detecting leaks from geological storage because they are unable to provide a sound basis for detecting shifts at unknown times. Information theoretic and Bayesian methods are preferable because they provide a statistically sound basis for detecting shifts at unknown times and are capable of handling non-Gaussian model errors.

Classical methods like the t-test suffer additionally because they provide poor control over the false alarm rate when errors are non-Gaussian. This problem is pronounced when few observations are available for detection. The Central Limit theorem may not apply, even for relatively large numbers of samples. When errors are not Gaussian, Bayesian methods should be used instead.

When errors are autocorrelated, standard methods like t-tests experience high false alarm rates because they tend to overestimate confidence in the sample mean. If additional systematic effects can be captured by improving the detrending model, it may be possible to reduce autocorrelation by improving the detrending model. In general, some degree of autocorrelation will remain due to missing and misrepresented physics in the model and (at minimum) nonzero instrumental response time. Therefore, when detecting leaks it will be important to use some method capable of accounting for autocorrelation. Estimation methods like the Yule-Walker method exist for adjusting uncertainty estimates in the presence of autocorrelation.

Forecast uncertainty can also be affected by choice of estimation method. If errors are Gaussian and the model is linear, standard methods like least squares apply. However, when few baseline samples are available and errors are not Gaussian, least squares can misrepresent and may underestimate uncertainty in the model parameters. The same problem can be encountered even when the errors are Gaussian if the model is nonlinear. Bayesian methods are capable of handling both scenarios; nonlinear least squares is less computationally intensive, however, and is applicable in the second case. The appropriate choice of estimation method will depend on the nature of the errors and the model being entertained. The consequence of using linear

least squares when the model is nonlinear or errors are non-Gaussian is the risk of increased false alarm rate.

9.2 CO_2 variability and models

Past studies of CO_2 variability and observations from the sites we have considered confirm the presence of systematic interannual, seasonal and diurnal trends.

Most interannual variability is captured assuming a constant fraction of emissions from fossil fuel combustion remains airborne. The growing trend is currently ~2 ppm, and appears to be growing with increasing rates of fossil fuel consumption. Relatively small deviations occur due to processes that even sophisticated models have difficulty explaining. Consistent interannual trends are present across sites.

Seasonal cycles are present, moreso in the Northern Hemisphere, which are due primarily to terrestrial photosynthesis and respiration. The magnitude of the seasonal cycle tends to grow larger north of the equator with increasing seasonal changes in sunlight and vegetation. The seasonal amplitude at Mauna Loa is about 5 ppm. The amplitude in Barrow, Alaska is approximately 15 ppm. Seasonality is minimal south of the equator, typically 1 ppm amplitude at most. Systematic changes in seasonal amplitude have been observed at Mauna Loa due to changes in terrestrial sinks and wind patterns.

Diurnal deviations from seasonal trends tend to be even larger in magnitude than the seasonal cycles themselves. Both local (~1 km) and regional scale (1 to 100+ km) processes affect diurnal variability. Diurnal trends therefore depend not only on the direct vicinity of the sensor but also upon the regional context. At Mauna Loa, local influences are minimal except occasionally at night when downslope winds bring CO_2 emitted from the volcano's summit. At Harvard Forest, both regional and local influences are important but regional biotic contributions tend to be bigger. Upwind combustion can also contribute up to 5 ppm enhancements in the winter. In Salt Lake City, diurnal variability is dominated by fossil fuel combustion, although respiration does contribute significantly in the summer. The scale of diurnal variability in Salt Lake is much greater than at Harvard Forest.

Several physically inspired, correlative models have been introduced for detrending atmospheric CO_2 data. Slope-intercept and emissions models are introduced to explain interannual trends. Solar radiation, dynamic leaf, and semi-annual harmonic terms are introduced to explain seasonal changes of biological origin and deviations from interannual trends whose origins are more complex.

Given several candidate models for CO_2, how should we choose which one to use? Clearly there is a need for a model that fits the data well. Methods like RSS can be used but do not provide clear stopping points because they only value goodness-of-fit: better fits can always be achieved by adding more terms. Information criteria (AIC, BIC, and DIC; [33, 31]) and Bayes Factors [95, 96] value fit but also avoid "overfitting the elephant" [97]. Information criteria are preferable for being easier to compute. We use AIC and DIC. AIC does not acknowledge uncertainty in the model parameters. DIC does acknowledge this uncertainty and represents a mix between

BIC and AIC. When the parameters are extremely uncertain, DIC is preferable from a Bayesian standpoint. Given the relatively small degree of uncertainty in our forward predictions, the difference between AIC and DIC is not substantial. Both give the same answer (model SRO) for Harvard Forest, which is where we expect the models proposed to perform best.

The seasonal models' inability to capture diurnal variability in the observations is the biggest contributor to uncertainty in future observations of atmospheric CO_2. Smoothing the observations by applying moving average filters with widths of at least $T = 1$ day can help reduce misfit errors. The biggest downside to using simple filters is that they cannot account for changes in the physical system. A better approach would be to improve the model by adding correlative terms linking diurnal fluctuations in CO_2 at the location of interest to changes in CO_2 at upwind sensors as well as to ancillary variables such as CO concentration. A worthwhile next step would be to build and screen new models of diurnal trends. The correlation structure of atmospheric CO_2 will need to be studied further to determine the typical spacing required for upwind sensors. This spatial aspect of the network design problem merits further attention.

9.3 The urban weekend weekday CO_2 shift

The ultimate goal for monitoring and verification is to estimate the magnitude of leakage. This requires inverting the shift in atmospheric CO_2 to obtain the flux distribution of the upwind source. In the absence of an actual leak, this can be done computationally. A different approach is taken here, which is to analyze the data from Harvard Forest and the University of Utah for real shifts tied to changes in urban emissions. The hypothesis is that increased urban activity during the workweek will lead to higher CO_2 concentrations on weekdays than weekends. The first question is whether this weekend weekday shift is detectable. If it is, the second question is what magnitude change in emissions (flux) led to the increase. The goal is to quantify what magnitude leak is detectable, given CO_2 data from Harvard Forest and Utah.

The Winter AM shift at Utah is +7.86 ppm[1]. The 95% confidence intervals[2] for the weekend and weekday residuals are -0.81 ± 1.35 ppm and 7.05 ± 0.96 ppm, respectively. A positive weekend weekday shift is clearly present in Salt Lake City, Utah.

The Winter AM shift at Harvard Forest is +1.13 ppm. The 95% confidence intervals for the weekend and weekday residuals are -1.59 ± 0.32 ppm and -0.46 ± 0.22 ppm, respectively, which leads us to believe a positive weekend weekday shift may also be present at Harvard Forest.

The Winter AM shift at Mauna Loa is +0.019 ppm. The 95% confidence intervals for the weekend and weekday residuals are -0.027 ± 0.036 ppm and -0.008 ± 0.024 ppm. The weekend weekday shift is not statistically significant, as expected given Mauna Loa observatory's remote location.

[1]Obtained as the difference in Mon-Fri and Sat-Sun means, 7.81 ppm $-(-0.81)$ ppm.

[2]These are given as mean $\pm$ 2$\times$(standard deviation).

The above discussion suggests the a shift is not detectable at Mauna Loa, and that a bigger shift is detectable at the urban University of Utah site than the rural Harvard Forest site. The signal to noise ratios in Table 7.1 support this trend. The signal to noise ratios of the weekend weekday shift are greatest (on winter mornings) for Utah (0.278), smaller for Harvard Forest (0.144), and near zero for Mauna Loa (0.027).

It is interesting to consider what would happen if all times were considered. The tendency observed at Harvard Forest and Utah is for signal to noise ratio to decrease. The reason this occurs is most likely that residuals insensitive to shifts in urban emissions are being included when the sample mean is calculated. The implication for leak detection is that by considering all times we might inadvertently dilute the shift being sought after. Clearly this will hinder our ability to detect a shift. This issue would not be mitigated by introducing models capable of explaining diurnal changes in CO_2. It is an issue of sensitivity. The sensitivity of the observations change as a function of time (of day), leading some times to be better candidates for detection than others. This issue is not one we have studied in depth and may merit further attention.

The residuals used to detect the weekend weekday CO_2 shift at Harvard and Utah failed the normality tests we applied at the 0.001 level of significance. A Bayesian detection algorithm based on Bayes factors was therefore conceived for testing.

The results from the algorithm suggest time to detection is sensitive to the distribution assumed for the residuals (empirical versus normal). Since the residuals are not Gaussian, the results based on empirical distributions should be trusted. The shift at Harvard was detectable at the 0.05 level of significance in 120-240 samples. As expected, the shift at Utah was detectable at the same level of significance in fewer sample (48-120), given the bigger signal to noise ratio at that site. If samples are recorded hourly, we must sample for at least 2 days at Utah and 5 days at Harvard Forest to determine whether the weekday CO_2 shift is present.

Comparison of Phoenix and Salt Lake City suggest the urban weekend weekday shift at the latter should be $\sim O(0.01)$ $MtCO_2km^{-2}yr^{-1}$. The biggest scale sequestration projects today like Sleipner inject $\sim O(1)$ $MtCO_2yr^{-1}$ belowground. Thus, it should be possible to detect a 1% leak rate[3] occurring over 1 km^2, simply by removing the seasonal signal from atmospheric CO_2 concentrations downwind of the source and testing the residuals for an upward mean shift.

Factors affecting the size leak that is detectable include the sensor's distance from the leak, leak area, and atmospheric stability. Detecting the shift will become difficult the farther the sensor is from the leak, as illustrated through the Harvard Forest example. Leakage at lower rate over bigger area will decrease the signal to noise ratio at the sensor, and may require improving models to explain higher frequency, diurnal effects for detection.

[3]Defined as 1% of the amount stored per year.

Chapter 10

Future directions

This thesis contributes a statistical framework for combining data, models, and detection algorithms to detect at a given site. The following represents a list of topics for future research that is needed to ensure monitoring and verification is available for geological sequestration:

- *Spatial variability.* This thesis has focused on temporal aspects of the leak detection problem. One of the key findings is that diurnal variability contributes most uncertainty in forecasts of CO_2 at sites subject to local influences. An easy way to deal with such variability would be to use multiple sensors. Since the goal is to detect perturbations above a baseline, additional sensors should be placed upwind of the sensor where the leak is to be detected. An obvious question is, how far away should they be placed? Addressing this will require studying sub-grid scale CO_2 variability, either through simulation of CO_2 concentrations or using measurements from the field.

- *Optimal network design.* Placing additional sensors in the field will help reduce detection limits, but improved performance will come at some extra cost. This is just one operational constraint for monitoring and verification. Others include safety and regulatory constraints. Finding the optimal design for a given location will require balancing the need to gather more information (for example, by adding sensors) against the operational constraints [19]. The following are some we think might be of future interest:

  - Is one design sufficient for most sites, or do alternative approaches need to be developed at different sites?
  - Can cheaper, less precise sensors be used?
  - Can ad-hoc wireless networks be used in combination with gas sensors for monitoring?

Appendix A

Predictive densities

A.1 Bayesian perspective

Predictive densities represent uncertainty in the model predictions. From a Bayesian perspective, predictive densities are obtained by marginalizing the joint density of the observations and unknowns over the unknowns. Thus, Bayesian predictive densities involve integrals like

$$f(Y) = \int f(Y|\theta)\pi(\theta)d\theta \tag{A.1}$$

The predictive density in (A.1) is called a *prior predictive density* because the conditional density of the observations is averaged over the prior knowledge of the unknowns. Alternative predictive densities can be constructed by averaging the conditional density of the observations over posterior density functions. Assuming the predictions $Y_{S'}$ are conditionally independent of a subset of the observations $Y_{S,obs}$, these alternative densities are defined by

$$f(Y_{S'}|Y_{S,obs}) = \int f(Y_{S'}|\theta)\pi(\theta|Y_{S,obs})d\theta \tag{A.2}$$

Table A.1 defines a few useful predictive densities with the help of (A.2).

Note that the predictive densities in (A.1) and (A.2) depend on the chosen model. They could be written $f(Y|\mathcal{M}_i)$ and $f(Y|Y_{obs}, \mathcal{M}_i)$. The meaning is the same, so the more explicit form will only be used as necessary to remove ambiguity.

A.2 Frequentist perspective

From the frequentist's perspective, θ is not random and should not be treated as such. As inconvenient as this perspective is, it does form the basis for many of the most common methods of inference and model selection. The analogy to (A.1) is achieved by pretending θ is known exactly, with some value θ_0. The prior density is then a Dirac delta function $\delta(\theta, \theta_0)$ and (A.1) simplifies to

$$f(Y) = f(Y|\theta_0) \tag{A.3}$$

Table A.1: Alternative forms of predictive densities

Type	Definition	Notes
Prior	$S = \emptyset$	The basis for the most basic form of Bayes factor. Not useful in practice because the prior $\pi(\theta)$ is often improper.
Posterior	$S = \{1, 2, ..., n\}$ and $S' \subseteq \{1, 2, ..., n\}$	Just use all the observations available.
Cross-validation	$S \subset \{1, 2, ..., n\}$ and $S' = \{1, 2, ..., n\} \setminus S$	The basis for *pseudo-Bayes factors.*
Forecasting	$S = \{1, 2, ..., n\}$ and $S' = \{n + 1\}$	Useful for forecasting applications.
Instrinsic	$S =$ minimal* and $S' = \{1, 2, ..., n\} \setminus S$	*S is a minimal set required to obtain a proper posterior $\pi(\theta \mid Y_{S,obs})$. The basis for *intrinsic Bayes factors.*

Thus, the predictive density is just the likelihood. Equation (A.3) forms the basis for classical approaches to model selection like confidence intervals, significance tests and likelihood ratio tests.

Appendix B

The Lomb Periodogram

The gaps that are so often present in CO_2 observations make it difficult to ascertain whether periodic signals are present. Most methods for testing autocorrelation (for example, the Breusch-Godfrey and confidence interval tests introduced in the thesis) assume the samples are spaced at equal time intervals Δt. Gaps make applying such tests difficult. The Lomb periodogram is capable of detecting harmonic content when the samples are unevenly distributed.

Lomb [98, 41] developed a method for estimating power spectra that is better equipped for detecting harmonic content in unevenly sampled series than FFT methods. The value of the Lomb normalized periodogram at a given angular frequency ω is defined by

$$P_N(\omega) = \frac{1}{2\hat{\sigma}^2} \left\{ \frac{\left[\sum_j (y_j - \hat{y}) \cos \omega \Delta t_j\right]}{\sum_j \cos^2 \omega \Delta t_j} + \frac{\left[\sum_j (y_j - \hat{y}) \sin \omega \Delta t_j\right]}{\sum_j \sin^2 \omega \Delta t_j} \right\} \tag{B.1}$$

where y_j is the value of the signal being examined at time t_j and where $\hat{y}$ and $\hat{\sigma}^2$ are the sample mean and variance based on samples $j = 1..n$. The idea is that differences from the sample mean that correlate well with a given harmonic lead to large values of power P_N at that frequency. Since the signals being considered are usually noisy, thresholds are often calculated above which harmonic content is said to be significant. The idea behind construction of these thresholds is just another an application of significance testing. The null hypothesis is that the samples y_j are $N(\bar{y}, \sigma^2)$. If the samples are somewhat evenly spaced (which is the case for most of the data we analyze, with the exception of those times when large gaps may be present), and the periodogram is generated for a large number of frequencies between 0 and the Nyquist critical frequency (which is effectively $1/2\Delta$ where $\Delta = (t_f - t_0)/n$ and t_0 and t_f are the first and last times considered), then the threshold is approximately

$$P_N^* \approx \ln \frac{n}{\alpha} \tag{B.2}$$

where α is the chosen significance level [41]. The above approximation improves as $alpha \to 0$.

The Lomb periodograms in this thesis were generated using code by [99]. The basis for this code is the algorithm outlined by Press [41], which computes (B.1).

Bibliography

[1] IPCC. *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, 2007.

[2] C.D. Keeling, T.P. Whorf, M. Wahlen, and J. van der Plichtt. Interannual extremes in the rate of rise of atmospheric carbon dioxide since 1980. *Nature*, 375:666–670, 1995.

[3] U. Siegenthaler, T.F. Stocker, E. Monnin, D. Luthi, J. Schwander, B. Stauffer, D. Raynaud, J-M Barnola, H. Fischer, V. Masson-Delmotte, and J. Jouzel. Stable carbon cycleclimate relationship during the late pleistocene. *Science*, 310:1313–1317, 2005.

[4] P. Forster, V. Ramaswamy, P. Artaxo, T. Berntsen, R. Betts, D.W. Fahey, J. Haywood, J. Lean, D.C. Lowe, G. Myhre, J. Nganga, R. Prinn, G. Raga, M. Schulz, and R. Van Dorland. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, chapter Changes in Atmospheric Constituents and in Radiative Forcing. Cambridge University Press, 2007.

[5] P.P. Tans. Recent global CO_2, November 2010.

[6] I.C. Prentice, G.D. Farquhar, M.J.R. Fasham, M.L. Goulden, M. Heimann, V.J. Jaramillo, H.S. Kheshgi, C. Le Quere, R.J. Scholes, and D.W.R. Wallace. *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, chapter The carbon cycle and atmospheric carbon dioxide, pages 184–238. Cambridge University Press, 2001.

[7] R.J. Francey, P.P. Tans, C.E. Allison, I.G. Enting, J.W.C. White, and M. Trolier. Changes in oceanic and terrestrial carbon uptake since 1982. *Nature*, 373:326–330, 1995.

[8] D.F. Ferretti, D.C. Lowe, R.J. Martin, and G.W. Brailsford. A new gc-irms technique for high precision, N_2O-free analysis of $\delta^{13}C$ and $\delta^{18}O$ in atmospheric CO_2 from small air samples. *Journal of Geophysical Research*, 105(D5):6709–6718, 2000.

[9] C.D. Keeling, A.F. Bollenbacher, and T.P. Whorf. *Trends: A Compendium of Data on Global Change*, chapter Monthly atmospheric $^{13}C/^{12}C$ isotopic ratios for 10 SIO stations. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN.

[10] R.F. Keeling, S.C. Piper, and M. Heimann. Global and hemispheric CO_2 sinks deduced from changes in atmospheric O_2 concentration. *Nature*, 381:218–221, 1996.

[11] R.F. Keeling and S.R. Shertz. Seasonal and interannual variations in atmospheric oxygen and implications for the global carbon cycle. *Nature*, 358:723–727, 1992.

[12] A.C. Manning and R.F. Keeling. Global oceanic and land biotic carbon sinks from the scripps atmospheric oxygen flask sampling network. *Tellus*, 58B:95–116, 2006.

[13] K.L. Denman, G. Brasseur, A. Chidthaisong, P. Ciais, P.M. Cox, R.E. Dickinson, D. Hauglustaine, C. Heinze, E. Holland, D. Jacob, U. Lohmann, S Ramachandran, P.L. da Silva Dias, S.C. Wofsy, and X. Zhang. *Couplings Between Changes in the Climate System and Biogeochemistry. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, 2007.

[14] U.S. Department of Energy. International Energy Outlook 2010. Technical Report DOE/EIA-0484(2010), Energy Information Administration, 2010.

[15] U.S. Department of Energy. Strategic plan. Technical Report DOE/PI-0005, U.S. Climate Change Technology Program, 2006.

[16] S. Pacala and R. Socolow. Stabilization wedges: Solving the climate problem for the next 50 years with current technologies. *Science*, 305:968–972, 2004.

[17] Massachusetts Institute of Technology. The future of coal. Technical report, 2007.

[18] IPCC. *IPCC Special Report on Carbon Dioxide Capture and Storage.* Cambridge University Press, 2005.

[19] C.J. Seto and G.J. McRae. Reducing risk in basin scale CO_2 sequestration: a framewok for integrated monitoring design. *Environmental Science and Technology*, 45(3):845–859, 2011.

[20] T. Ide, S.J. Friedmann, and H. Herzog. CO_2 leakage through existing wells: current technology and regulations. In *8th International Conference on Greenhouse Gas Control Technologies*, 2006.

[21] G.W. Scherer, M.A. Celia, J-H Prevost, S. Bachu, R. Bruant, A. Duguid, R. Fuller, S.E. Gasda, M. Radonjic, and W. Vichit-Vadakan. *Carbon Dioxide Capture for Storage in Deep Geologic Formations Results from the CO_2 Capture Project: Geologic Storage of Carbon Dioxide with Monitoring and Verification*, volume 2, chapter Leakage of CO_2 through abandoned wells: role of corrosion of cement. Elsevier, 2005.

[22] Y.-M. Yang. *Statistical Methods for Integrating Multiple CO_2 Leak Detection Techniques at Geologic Sequestration Sites.* PhD thesis, Carnegie Mellon University, 2011.

[23] T. Andersen, J. Carstensen, E. Hernández-Garcia, and C.M. Duarte. Ecological thresholds and regime shifts: approaches to identification. *Trends in Ecology and Evolution*, 24(1):49–57, 2009.

[24] S.N. Rodionov. A brief overview of the regime shift detection methods. In V. Velikova and N. Chipev, editors, *Large-Scale Disturbances (Regime Shifts) and Recovery in Aquatic Ecosystems: Challenges for Management Toward Sustainability*, pages 17–24, 2005.

[25] J.D. Testerman. A statistical reservoir-zonation technique. *Journal of Petroleum Technology*, pages 889–893, August 1962.

[26] D. Gill. Application of a statistical zonation method to reservoir evaluation and digitized-log analysis. *The American Association of Petroleum Geologists Bulletin*, 54(5):719–729, May 1970.

[27] D.M. Hawkins. Fitting multiple change-point models to data. *Computational statistics and data analysis*, 37:323–341, 2001.

[28] M. Lavielle. Optimal segmentation of random processes. *IEEE Transactions on Signal Processing*, 46(5):1365–1373, 1998.

[29] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[30] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[31] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64:583–639, 2002.

[32] E.J. Hannan and B.G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series B*, 41:190–195, 1979.

[33] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference.* Springer-Verlag, 2002.

[34] K.P. Burnham. Multimodel inference: Understanding AIC and BIC in model selection. In *Amsterdam Workshop on Model Selection*, 2004.

[35] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice.* Chapman & Hall/CRC, 1996.

[36] D. Gamerman and H.F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman & Hall/CRC, 2006.

[37] R.B. D'Agostino and M.A. Stephens. *Goodness-of-Fit Techniques.* Marcel Dekker, 1986.

[38] C.M. Jarque and A.K. Bera. A test for normality of observations and regression residuals. *International Statistical Review*, 55(2):163–172, 1987.

[39] G. A.F. Seber and C. J. Wild. *Nonlinear Regression.* John Wiley & Sons, Inc., 2003.

[40] A. Sen and M. Srivastava. *Regression Analysis: Theory, Methods and Applications.* Springer-Verlag, 1990.

[41] W.H. Press. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, 2007.

[42] Q. Liu, L. Gu, R.E. Dickinson, Y. Tian, L. Zhou, and W.M. Post. Assimilation of satellite reflectance data into a dynamical leaf model to infer seasonally varying leaf areas for climate and carbon models. *Journal of Geophysical Research*, 113(D19113), 2008.

[43] D.F. Nicholls and A.R. Pagan. Specification of the disturbance for efficient estimation–an extended analysis. *Econometrica*, 45(1):211–217, 1977.

[44] R. Davidson and J.G. MacKinnon. *Estimation and Inference in Econometrics.* Oxford University Press, 1993.

[45] A.K. Bera, M.L. Higgins, and S. Lee. Interaction between autocorrelation and conditional heteroscedasticity: a random-coefficient approach. *Journal of Business & Economic Statistics*, 10(2):133–142, 1992.

[46] S.T. Sathe and H.D. Vinod. Bounds on the variance of regression coefficients due to heteroscedastic or autoregressive errors. *Econometrica*, 42(2):333–340, 1974.

[47] J. Johnston and J. DiNardo. *Econometric Methods.* McGraw-Hill, 1997.

[48] T.S. Breusch. Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, 17(31):334–355, 1978.

[49] T.S. Breusch and A.R. Pagan. The Lagrange Multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, 47(1):239–253, 1980.

[50] L.G. Godfrey. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, 46(6):1293–1301, 1978.

[51] Hirotugu Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247, 1969.

[52] A.R. Gallant and J.J. Goebel. Nonlinear regression with autocorrelated errors. *Journal of the American Statistical Association*, 71(356):961–967, 1976.

[53] W.D. Komhyr, T.B. Harris, and L.S. Waterman. Atmospheric carbon dioxide at Mauna Loa Observatory, 1. NOAA global monitoring for climatic change measurements with a nondispersive infrared analyzer, 1974-1985. *Journal of Geophysical Research*, 94(D6):8533–8547, 1989.

[54] H.D. Vinod. Effects of ARMA errors on the significance tests for regression coefficients. *Journal of the American Statistical Association*, 71(356):929–933, 1976.

[55] J.M. Zobitz, J.P. Keener, H. Schnyder, and D.R. Bowling. Sensitivity analysis and quantification of uncertainty for isotopic mixing relationships in carbon cycle research. *Agricultural and Forest Meteorology*, 136:56–75, 2006.

[56] J.W. Moore and B.X. Semmens. Incorporating uncertainty and prior information into stable isotope mixing models. *Ecology Letters*, 11:470–480, 2008.

[57] J.B. West, G.J. Bowen, T.E. Cerling, and J.R. Ehleringer. Stable isotopes as one of Natures ecological recorders. *Trends in Ecology and Evolution*, 21(7):408–414, 2006.

[58] S.E. Bush, D.E. Pataki, and J.R. Ehleringer. Sources of variation in $\delta^{13}C$ of fossil fuel emissions in Salt Lake City, USA. *Applied Geochemistry*, 22:715–723, 2007.

[59] D.J. Hofmann, J.H. Butler, and P.P. Tans. A new look at atmospheric carbon dioxide. *Atmospheric Environment*, 43:2084–2086, 2009.

[60] C.D. Keeling, R.B. Bacastow, A.F Carter, S.C. Piper, T.P. Whorf, M. Heimann, W.G. Mook, and H. Roeloffzen. *Aspects of Climate Variability in the Pacific and Western Americas*, chapter A three-dimensional model of atmospheric CO_2 transport based on observed winds: 1. Analysis of observational data, pages 165–236. Number 55 in Geophysical Monograph. American Geophysical Union, Washington, D.C., 1989.

[61] R.J. Francey, C.M. Trudinger, M. Van der Schoot, P.B. Krummel, L.P. Steele, and R.L. Langenfelds. Differences between trends in atmospheric CO_2 and the reported trends in anthropogenic CO_2 emissions. *Tellus*, 62B:316–328, 2010.

[62] T.J. Conway and P.P. Tans. Development of the CO_2 latitude gradient in recent decades. *Global biogeochemical cycles*, 13(4):821–826, 1999.

[63] W. Buermann, B.R. Lintner, C.D. Koven, A. Angert, J.E. Pinzon, C.J. Tucker, and I.Y. Fung. The changing carbon cycle at Mauna Loa Observatory. *PNAS*, 104(11):4249–4254, March 2007.

[64] S.C. Wofsy, M.L. Goulden, J.W. Munger, S.-M. Fan, P.S. Bakwin, B.C. Daube, S.L. Bassow, and F.A. Bazzaz. Net exchange of CO_2 in a mid-latitude forest. *Science*, 260(5112):1314–1317, 1993.

[65] R.M. Law, W. Peters, C. Roedenbeck, C. Aulagnier, I. Baker, D.J. Bergmann, P. Bousquet, J. Brandt, L. Bruhwiler, P.J. Cameron-Smith, J.H. Christensen, F. Delage, A.S. Denning, S. Fan, C. Geels, S. Houweling, R. Imasu, U. Karstens, S. R. Kawa, J. Kleist, M.C. Krol, S.-J. Lin, R. Lokupitiya, T. Maki, S. Maksyutov, Y. Niwa, R. Onishi, N. Parazoo, P.K. Patra, G. Pieterse, L. Rivier, M. Satoh, S. Serrar, S. Taguchi, M. Takigawa, R. Vautard, A.T. Vermeulen, and Z. Zhu. Transcom model simulations of hourly atmospheric CO_2: Experimental overview and diurnal cycle results for 2002. *Global Biogeochemical Cycles*, 22(GB3009), August 2008.

[66] R.M. Law, L.P. Steele, P.B. Krummel, and W. Zahorowski. Synoptic variations in atmospheric CO_2 at Cape Grimm: a model intercomparison. *Tellus*, 62B:810–820, 2010.

[67] R. Ahmadov, C. Gerbig, R. Kretschmer, S.Koerner, C.Roedenbeck, P. Bousquet, and M. Ramonet. Comparing high resolution WRF-VPRM simulations and two global CO_2 transport models with coastal tower measurements of CO_2. *Biogeosciences*, 6:807–817, 2009.

[68] C.D. Keeling, S.C. Piper, R.B. Bacastow, M. Wahlen, T.P. Whorf, M. Heimann, and H.A. Meijer. Exchanges of atmospheric CO_2 and $^{13}CO_2$ with the terrestrial biosphere and oceans from 1978 to 2000. I. Global aspects. SIO Reference Series No. 01-06, Scripps Institution of Oceanography, San Diego, 2001. Last updated 21-Apr-2010.

[69] C.D. Keeling, R.B. Bacastow, A.E. Bainbridge, C.A. Ekdahl, P.R. Guenther, and L.S. Waterman. Atmosheric carbon dioxide variations at Mauna Loa Observatory, Hawaii. *Tellus*, 28(6):538–551, 1976.

[70] R.B. Bacastow, C.D. Keeling, and T.P. Whorf. Seasonal amplitude increase in atmospheric concentration at Mauna Loa, Hawaii, 1959-1982. *Journal of Geophysical Research*, 90(D6):10529–10540, 1985.

[71] B.G. Mendonca. Local wind circulation on the slopes of Mauna Loa. *Journal of Applied Meteorology*, 8:533–541, 1969.

[72] S. Ryan. The wind field around Mauna Loa derived from surface and balloon observations. *Journal of Geophysical Research*, 102(D9):10711–10725, 1997.

[73] K.W. Thoning, P.P. Tans, and W.D. Komhyr. Atmospheric carbon dioxide at Mauna Loa Observatory 2. Analysis of the NOAA GMCC data, 1974-1985. *Journal of Geophysical Research*, 94(D6):8549–8565, 1989.

[74] S. Ryan. Estimating volcanic CO_2 emission rates from atmospheric measurements on the slope of Mauna Loa. *Chemical Geology*, 177:201–211, 2001.

[75] S. Urbanski, C. Barford, S. Wofsy, C. Kucharik, E. Pyle, J. Budney, K. McKain, D. Fitzjarrald, M. Czikowsky, and J.W. Munger. Factors controlling CO_2 exchange on timescales from hourly to decadal at Harvard Forest. *Journal of Geophysical Research*, 112(G02020), 2007.

[76] C.C. Barford, S.C. Wofsy, M.L. Goulden, J.W. Munger, E.H. Pyle, S.P. Urbanski, L.Hutyra, S.R. Saleska, D. Fitzjarrald, and K. Moore. Factors controlling long- and short-term sequestration of atmospheric CO_2 in a mid-latitude forest. *Science*, 294:1688–1691, 2001.

[77] M.J. Potosnak, S.C. Wofsy, A.S. Denning, T.J. Conway, J.W. Munger, and D.H. Barnes. Influence of biotic exchange and combustion sources on atmospheric concentrations in new england from observations at a forest flux tower. *Journal of Geophysical Research*, 104(D8):9561–9569, 1999.

[78] D.H. Barnes, S.C. Wofsy, B.P. Fehlau, E.W. Gottlieb, J.W. Elkins, G.S. Dutton, and S.A. Montzka. Urban/industrial pollution for the New York City–Washington, D. C., corridor, 1996–1998: 1. Providing independent verification of CO and PCE emissions inventories. *Journal of Geophysical Research*, 108(D6), 2003.

[79] D.E. Pataki, D.R. Bowling, J.R. Ehleringer, and J.M. Zobitz. High resolution atmospheric monitoring of urban carbon dioxide sources. *Geophysical Research Letters*, 33(L03813), 2006.

[80] D.E. Pataki, D.R. Bowling, and J.R. Ehleringer. The seasonal cycle of carbon dioxide and its isotopic composition in an urban atmosphere: anthropogenic and biogenic effects. *Journal of Geophysical Research*, 108:4735, 2003.

[81] D.E. Pataki, T. Xu, Y.Q. Luo, and J.R. Ehleringer. Inferring biogenic and anthropogenic carbon dioxide sources across an urban to rural gradient. *Oecologia*, 152:307–322, 2007.

[82] D.E. Pataki, D.R. Bowling, and J.R. Ehleringer. Seasonal cycle of carbon dioxide and its isotopic composition in an urban atmosphere: Anthropogenic and biogenic effects. *Journal of Geophysical Research*, 108(D23), 2003.

[83] D.E. Pataki, B.J. Tyler, R.E. Peterson, A.P. Nair, W.J. Steenburgh, and E.R. Pardyjak. Can carbon dioxide be used as a tracer of urban atmospheric transport? *Journal of Geophysical Research*, 110(D15102), 2005.

[84] R.E. Dickinson, Y. Tian, Q. Liu, and L. Zhou. Dynamics of leaf area for climate and weather models. *Journal of Geophysical Research*, 113(D16115), 2008.

[85] K.A. Masarie and P.P. Tans. Extension and integration of atmospheric carbon dioxide data into a globally consistent measurement record. *Journal of Geophysical Research*, 100(D6):11593–11610, 1995.

[86] T. Nakazawa, M. Ishizawa, K. Higuchi, and N.B.A. Trivett. Two curve fitting methods applied to CO_2 flask data. *Environmetrics*, 8:197–218, 1997.

[87] C.D. Nevison, N.M. Mahowald, S.C. Doney, I.D. Lima, G.R. van der Werf, J.T. Randerson, D.F. Baker, P. Kasibhatla, and G.A. McKinley. Contribution of ocean, fossil fuel, land biosphere, and biomass burning carbon fluxes to seasonal and interannual variability in atmospheric CO_2. *Journal of Geophysical Research*, 113:1–21, 2008.

[88] K. McKain and S. Wofsy. Urban CO_2 modeling project (Phase II): Final report to Scitor Corporation. April 2010.

[89] C.D. Idso, S.B. Idso, and R.C. Balling Jr. An intensive two-week study of an urban CO_2 dome in Phoenix, Arizona, USA. *Atmospheric Environment*, 35:995–1000, 2001.

[90] R.S. Cerveny and K.J. Coakley. A weekly cycle in atmospheric carbon dioxide. *Geophysical Research Letters*, 29(2), 2002.

[91] S.B. Idso, C.D. Idso, and R.C. Balling Jr. Seasonal and diurnal variations of near-surface atmospheric CO_2 concentration within a residential sector of the urban CO_2 dome of Phoenix, AZ, USA. *Atmospheric Environment*, 36:1655–1660, 2002.

[92] B. Koerner and J. Klopatek. Anthropogenic and natural CO_2 emission sources in an arid urban environment. *Environmental Pollution*, 116:S45–S51, 2002.

[93] U.S. Census Bureau. Metropolitan area population estimates for July 1, 1999 and population change for April 1, 1990 to July 1, 1999 (includes April 1, 1990 population estimates base), October 2000. MA-99-1.

[94] U.S. Census Bureau. Metropolitan and micropolitan statistical area estimates for April 1, 2000 to July 1, 2009, July 2009. CBSA-EST2009-01.

[95] R.E. Kass and A.E. Raftery. Bayes factors. *J. Amer. Stat. Assoc.*, 90(430):773–795, 1995.

[96] A.E. Gelfand and D.K. Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society Series B*, 56(3):501–514, 1994.

[97] J. Wei. Least squares fitting of an elephant. *CHEMTECH*, 5(2):128–129, 1975.

[98] N.R. Lomb. Least-squares frequency analysis of unequally-spaced data. *Astrophysics and Space Science*, 39:447–462, 1976.

[99] D. Savransky. Lomb (lomb-scargle) periodogram. Matlab File Exchange, May 2008.