

Machine Learning(SME3006) : Homework 5

CO₂ 농도 예측을 위한 BLR 및 GP 비교 분석

Jun Hyeong Doh

Department of Smart Mobility
Engineering

Inha University

November 2025

Chapter 1

서론

1.1 과제 목표 및 개요

대기 중 이산화탄소(CO_2) 농도는 지구 기후 변화를 이해하고 미래 환경 정책을 수립하는데 가장 중요한 지표 중 하나이다. NOAA(미국 해양대기청)에서 제공하는 월별 평균 CO_2 데이터는 명확한 장기 상승 추세와 주기적인 계절 변동성을 포함하는 시계열 데이터이다. 이러한 비선형성과 복잡성을 가진 시계열 데이터를 예측할 때, 단순히 예측 평균만을 제공하는 기존의 모델들은 예측의 신뢰도나 위험을 정량화할 수 없다는 한계를 갖는다.

본 보고서는 이러한 한계를 극복하고 예측의 불확실성까지 정량화하는 확률론적 회귀 방법을 탐구하는 것을 목적으로 한다. 특히 베이지안 선형 회귀(Bayesian Linear Regression, BLR)와 가우시안 프로세스(Gaussian Process, GP)라는 두 가지 베이지안 기법을 사용하여 CO_2 데이터를 분석한다.

베이지안 모델은 모델 파라미터 또는 함수 자체에 확률 분포를 부여하여, 예측 결과에 대한 불확실성(Uncertainty)를 통계적으로 표현한다. 이는 예측 오차에 내재된 Aleatoric Uncertainty(데이터 내재적 불확실성)과 데이터 부족에서 오는 Epistemic Uncertainty(모델 지식 부족)을 쉽게 분리하여 분석할 수 있게 한다.

본 과제의 구체적인 목표는 다음과 같다.

- **예측 성능 평가(2016-2025)** : 1958년부터 2015년까지의 데이터를 훈련하여 두 모델의 예측 정확도를 비교하고, 실제 관측값 대비 불확실성의 신뢰도를 평가한다.

- **미래 CO₂ 수준 예측(-2040년)** : 전체 데이터를 활용하여 모델을 재훈련하고, 2040년 까지의 CO₂ 수준을 예측 평균과 불확실성 대역으로 제시하여 장기 예측의 신뢰도와 환경적 의미를 논의한다.

1.2 사용 데이터

본 보고서에서는 미국 해양대기청(NOAA)에서 제공하는 마우나 로아(Mauna Loa) 관측소의 월평균 CO₂ 농도 데이터를 사용한다. 이 데이터는 1958년부터 현재까지 장기간에 걸쳐 일관되게 측정된 자료로, 전 지구적 CO₂ 농도 변화의 기준점으로 주로 쓰이고 있다.

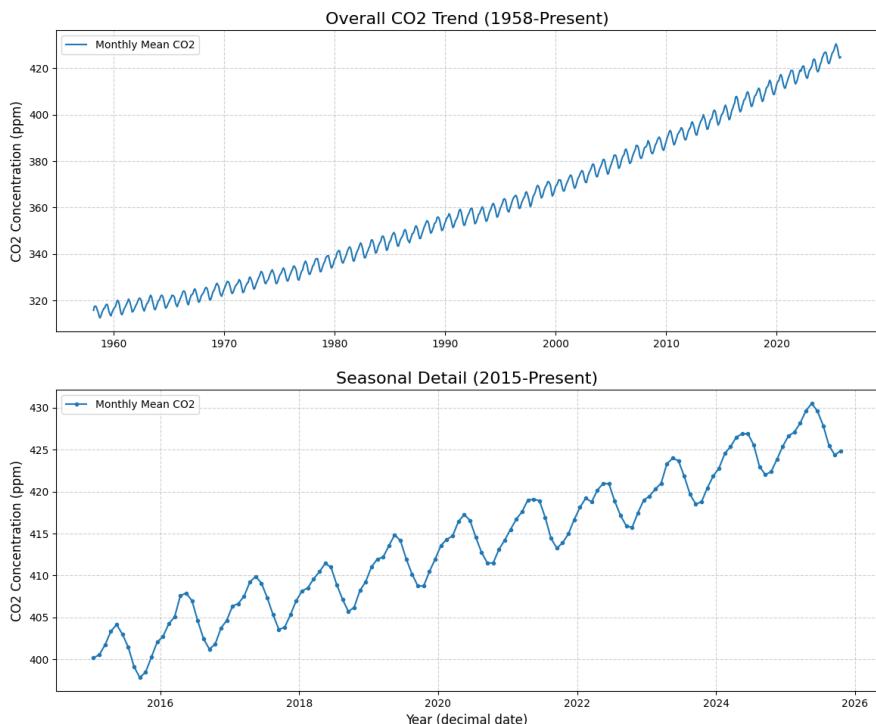


Figure 1.1: Mauna Loa CO₂ 농도의 장기적 추세 및 계절적 변동성

Figure 1.1와 같이 데이터는 명확한 장기적 상승 추세와 함께, 계절적 요인에 의한 주기적인 변동성을 특징으로 하고 있다. 확률적 모델을 통해 이러한 복잡한 시계열 패턴을 학습하고 미래를 예측하는 것이 본 보고서의 핵심이다.

Chapter 2

이론적 배경

일반적인 회귀 모델(Least Square 등)은 데이터에 가장 잘 맞는 단일 함수 $y = f(x)$ 를 찾는 것을 목표로 한다. 하지만 이러한 모델은 예측이 얼마나 확실한지에 대한 정보를 제공하지 않는다. 본 보고서에서는 데이터 예측의 불확실성 정량화(Uncertainty Quantification, UQ)를 핵심 목표로 한다.

불확실성을 모델링하기 위해, 본 보고서에서는 대표적인 두 가지 확률적 회귀 모델인 베이즈 선형 회귀(BLR)과 가우시안 프로세스(GP)를 사용하고자 한다.

2.1 베이즈 선형 회귀(Bayesian Linear Regression, BLR)

선형 회귀 모델은 $y = \phi(\mathbf{x})^T \mathbf{w} + \epsilon$ 로 표현되며, 여기서 $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ 은 관측 노이즈를 의미한다. $\phi(\mathbf{x})$ 는 입력 \mathbf{x} 를 특징 공간으로 매핑하는 기저 함수 벡터이다. 일반적인 선형 회귀는 \mathbf{w} 의 단일 최적해를 찾는 반면, 베이즈 선형 회귀(BLR)는 \mathbf{w} 를 고정된 값이 아닌 확률 변수로 추론한다.

- **사전 확률(Prior):** 우선 \mathbf{w} 에 대한 사전 가정을 정의한다. 보통 평균이 0인 가우시안 분포 $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \sigma_w^2 \mathbf{I})$ 로 정의한다.
- **가능도(Likelihood):** 훈련 데이터 $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}$ 가 주어졌을 때, \mathbf{w} 가 이 데이터들을 얼마나 잘 설명하는지에 대한 우도 $p(\mathcal{D} | \mathbf{w})$ 를 계산한다.
- **사후 확률(Posterior):** 베이즈 정리(Bayes' Theorem)을 사용해, 데이터를 관찰한 후

업데이트된 \mathbf{w} 의 사후 확률 분포 $p(\mathbf{w}|\mathcal{D})$ 를 구한다.

$$p(\mathbf{w}|X, y, \sigma^2) \propto p(y|X, \mathbf{w}, \sigma^2)p(\mathbf{w}) \quad (2.1)$$

가우시안 분포를 서로 곱해도 가우시안 분포이므로, 이 사후 확률 또한 가우시안 분포가 된다.

- **예측 분포(Predictive Distribution):** 새로운 입력 \mathbf{x}_* 에 대한 예측 y_* 는, 훈련 데이터 \mathcal{D} 를 학습한 \mathbf{w} 의 사후 확률 분포 $p(\mathbf{w}|\mathcal{D})$ 를 사용하여 계산한다. 이는 \mathbf{w} 의 모든 가능한 값에 대해 적분하여 평균을 내는 과정이며, 다음과 같은 수식으로 표현된다:

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \int p(y_*|\mathbf{x}_*, \mathbf{w})p(\mathbf{w}|\mathcal{D})d\mathbf{w} \quad (2.2)$$

2.2 가우시안 프로세스(Gaussian Process, GP)

가우시안 프로세스(GP)는 특정 파라미터(\mathbf{w})를 모델링하는 대신, 함수 자체에 대한 확률 분포를 직접 모델링하는 접근 방식이다.

GP는 평균 함수 $m(\mathbf{x})$ (보통 0으로 설정)과 공분산 함수(커널) $k(\mathbf{x}, \mathbf{x}')$ 에 의해 정의된다. 커널 함수는 두 입력 \mathbf{x} 와 \mathbf{x}' 에서의 함수 값 $f(\mathbf{x})$ 와 $f(\mathbf{x}')$ 의 공분산을 정의하며, 함수의 속성(부드러움, 주기성 등)을 결정한다.

훈련 데이터 \mathcal{D} 가 주어지면, GP는 사전 분포(Prior)를 데이터에 적용하여 사후 분포(Posterior)를 얻는다. 새로운 입력 \mathbf{x}_* 에 대한 예측 또한 예측 분포 $p(f_*|\mathbf{x}_*, \mathcal{D})$ 를 통해 이루어지며, 이는 가우시안 분포 $\mathcal{N}(f_*|\mu_*, \sigma_*^2)$ 가 된다.

GP의 가장 큰 장점은 커널 함수를 유연하게 조합할 수 있다는 점이다. 예를 들어, CO₂ 데이터의 복잡한 패턴을 모델링하기 위해 장기 추세를 위한 커널과 계절성을 위한 주기 커널을 조합하여 사용할 수 있다. 이는 기저 함수를 수동으로 설계할 필요 없이 데이터로부터 복잡한 구조를 학습할 수 있게 한다.

커널의 특징 상 다양한 커널들을 새로 조합할 수 있고, 그 커널 함수가 만들어내는 Prior에 맞춰서 우리가 풀고자 하는 문제에 대한 예측 모델을 가정할 수 있다. Figure 2.1에서와 같이, 데이터를 선형적으로 Fitting하는 것이 아니라, 수많은 가는 파란색 선이 Function의

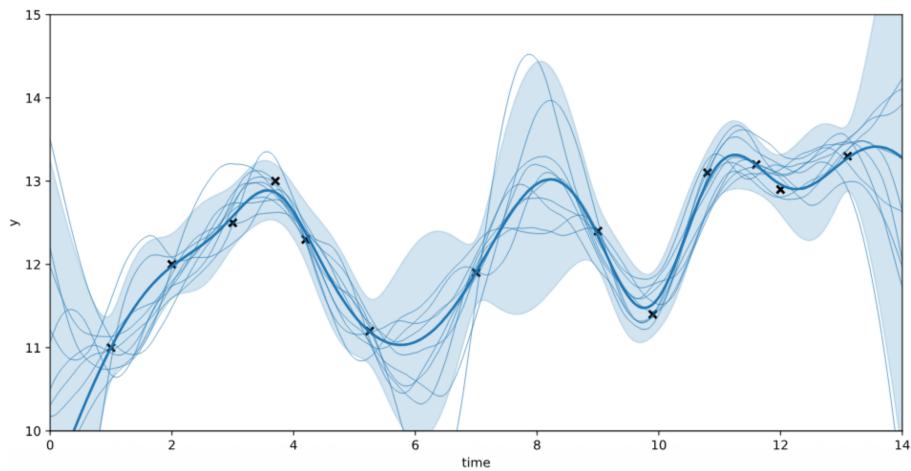


Figure 2.1: Gaussian Process 결과 예시

후보가 되어 그에 따라 비선형적으로 Fitting^o 된다. 데이터가 많은 구간에서는 Function들이 유사하게 나타나면서 불확실성이 낮게 나타나고(파란색 음영 부분이 좁음), 데이터가 없는 구간에서는 Function들이 다양하게 나타나면서 불확실성이 크게 나타난다(파란색 음영 부분이 넓음). 이렇듯 Gaussian Process를 통해 불확실성을 알아낼 수 있다.

Chapter 3

예측 성능 평가

본 챕터는 과제 1의 요구사항에 따라, 1958년부터 2015년까지의 데이터를 Training Set로 사용하여 2016년부터 2025년까지의 CO₂ 농도를 예측하고, 그 성능을 실제 관측값과 비교 분석한다.

3.1 베이즈 선형 회귀(Bayesian Linear Regression, BLR)

본 절에서는 2장에서 설명한 이론을 바탕으로 베이즈 선형 회귀(BLR) 모델을 구현한다. 아래 Figure 3.1과 같이, 입력 x 가 기저 함수 $\phi(x)$ 를 통해 변환되고, 훈련 데이터(X, y)와 사전 확률 $P(w)$ 을 결합하여 사후 확률 $P(w|\mathcal{D})$ 가 계산된다. 이 사후 확률은 새로운 입력 $\Phi(x_*)$ 과 결합하여 최종 예측 분포를 도출한다.

3.1.1 1차 선형 기저 함수 및 한계

BLR 모델의 가장 기본적인 형태는 렉쳐 노트의 예시와 같이 1차 선형 함수 $\phi(x) = [1, x]$ ($M=2$)를 기저 함수로 사용하는 것이다.

모델 학습 및 하이퍼파라미터: 기저 함수와 훈련 데이터를 사용하여 모델 파라미터 w 의 사후 확률 $p(w|\mathcal{D}) = \mathcal{N}(w|\mu, \Sigma)$ 을 추정하였다. 사후 확률의 평균 μ 와 공분산 Σ 는 렉쳐 노트의 닫힌 형태(closed-form) 공식을 코드로 구현하여 다음과 같이 직접 계산하였다:

$$\Sigma = (\sigma_w^{-2}\mathbf{I} + \sigma^{-2}\mathbf{X}^T\mathbf{X})^{-1} \quad (3.1)$$

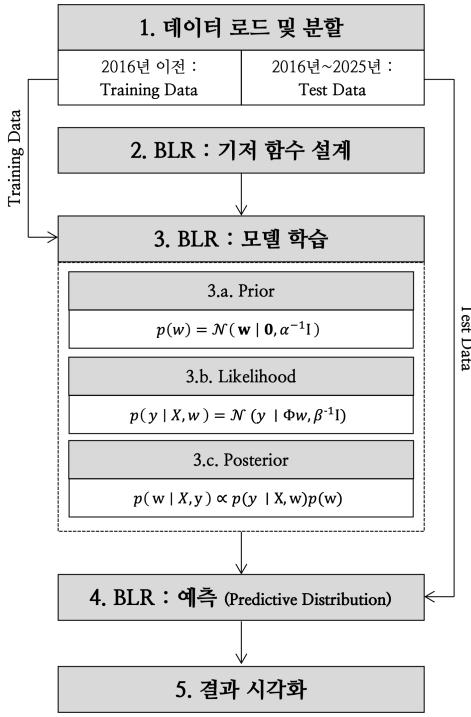


Figure 3.1: 베이즈 선형 회귀 순서도

$$\boldsymbol{\mu} = \sigma^{-2} \boldsymbol{\Sigma} \mathbf{X}^T \mathbf{y} \quad (3.2)$$

(여기서 \mathbf{X} 는 Φ_{train} , \mathbf{y} 는 y_{train} 을 의미한다.)

이때, 위 공식에 사용된 하이퍼파라미터는 다음과 같이 설정하였다:

- 파라미터 \mathbf{w} 의 사전 분산: $\sigma_w^2 = 100$
- 데이터 노이즈 분산: $\sigma^2 = 2^2 = 4$

사전 분산(σ_w^2)을 100으로 크게 설정한 것은, 파라미터에 대한 사전 정보를 약하게 부여하여 데이터가 \mathbf{w} 를 결정하도록 하기 위함이다.

노이즈 분산(σ^2)은 Figure 1.1의 계절적 변동폭을 고려하여 약 2 ppm의 표준편차를 가질 것으로 가정한 값이다. 이는 데이터의 계절적 진폭이 약 6~7 ppm인 점과, 모델이 포착하지 못할 미세한 기후 변동성이 약 1~2 ppm 존재할 것이라는 가정을 기저에 두고 있다. 또한 σ 를 다소 보수적으로 설정함으로써 기저 함수의 구조적 한계로 인한 모델 불일치를 불확실성

영역에 충분히 포함 시키고자 하였다.

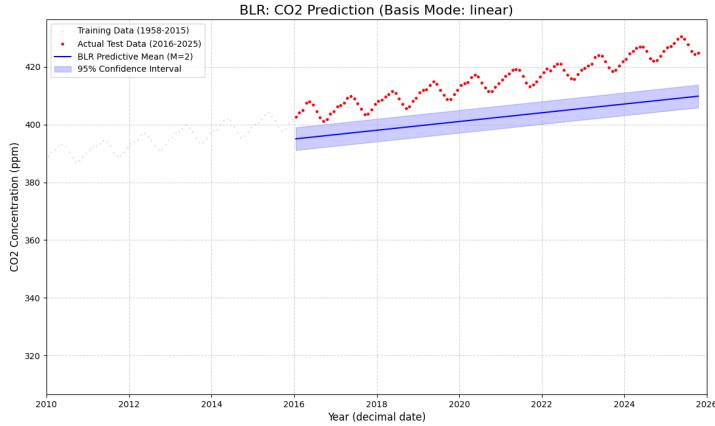


Figure 3.2: 1차 선형 기저 함수를 이용한 베이즈 선형 회귀 결과

이 모델을 1958-2015년 훈련 데이터에 적용하여 2016-2025년을 예측한 결과는 Figure 3.2과 같다. 예상대로, 예측 결과는 실제 데이터의 복잡한 패턴을 전혀 반영하지 못했다. 이 모델은 두 가지 명확한 한계를 보였다.

- **한계 1 (추세):** 직선 추세는 데이터의 장기적인 비선형 추세(CO_2 농도가 시간이 지남 수록 가속하는 추세)를 따라잡지 못해, 시간이 갈수록 오차가 크게 발생한다.
- **한계 2 (계절성):** 모델이 계절에 따른 주기적인 변동을 전혀 학습하지 못한다.

3.1.2 비선형 기저 함수 설계

앞선 1차 모델의 한계를 극복하기 위해, CO_2 데이터의 특징을 명시적으로 모델링할 수 있는 비선형 기저 함수를 도입하였다.

우선 전체적인 직선 추세를 반영하기 위해 식 3.3의 복합 기저 함수를 구성하였다. 여기서 x_{norm} 은 수치적 안정성을 위해 x 에서 훈련 데이터의 평균을 뺀 값이다.

$$\phi(x) = [1, x_{norm}, x_{norm}^2] \quad (3.3)$$

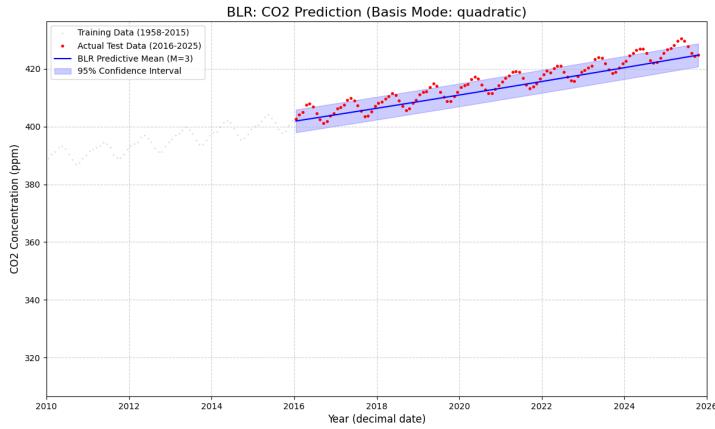


Figure 3.3: 3개의 기저 함수를 이용한 베이즈 선형 회귀 결과

위 식 3.3의 복합 기저 함수를 이용한 예측 분포는 위 Figure 3.3와 같다. 연도별 전체적인 추세는 잘 추종하지만, 계절에 따른 주기적인 변동을 학습하지 못했다.

앞선 $M=3$ 모델의 한계는 단순한 2차 함수로는 계절성을 표현할 수 없다는 점이었다. 이에 더해, 실제 CO_2 데이터의 계절적 변동은 완벽한 사인파가 아니라, 상승기와 하강기의 형태가 비대칭적인 특징을 갖는다.

최종적으로 본 보고서에서는 이러한 데이터의 복합적인 특성을 반영하기 위해 다음과 같은 논리로 기저 함수를 확장 설계하였다.

설계한 복합 기저 함수는 7차원이며, 다음과 같이 정의했다:

$$\phi(x) = [1, x_{norm}, x_{norm}^2, \sin(2\pi x), \cos(2\pi x), \sin(4\pi x), \cos(4\pi x)] \quad (3.4)$$

단순히 $\sin(2\pi x)$ 만 이용하는 경우, 데이터의 계절적 주기가 정확히 0 시점(1월 1일)에 0으로 시작해야 한다는 강한 조건이 생긴다. 하지만 실제 CO_2 의 농도 피크는 5월 경으로 보인다. 따라서 \sin, \cos 선형 결합을 이용해 파동의 위상과 진폭을 자유롭게 조절하고자 하였다.

또한, 데이터의 시간 단위 x 가 년 단위이므로, 1년 주기를 갖는 파동의 각진동수 $\omega = 2\pi f = 2\pi \times 1 = 2\pi$ 이므로, 이를 기본 주파수로 하였다.

마지막으로 CO_2 농도의 계절적 변동이 완벽한 정현파가 아니라는 점에 대응하기 위해, 기본

주파수(2π)에 2배의 주파수를 사용하는 2차 고조파(4π)를 추가하여 파형의 찌그러짐이나 뾰족함을 극사화 하도록 하였다.

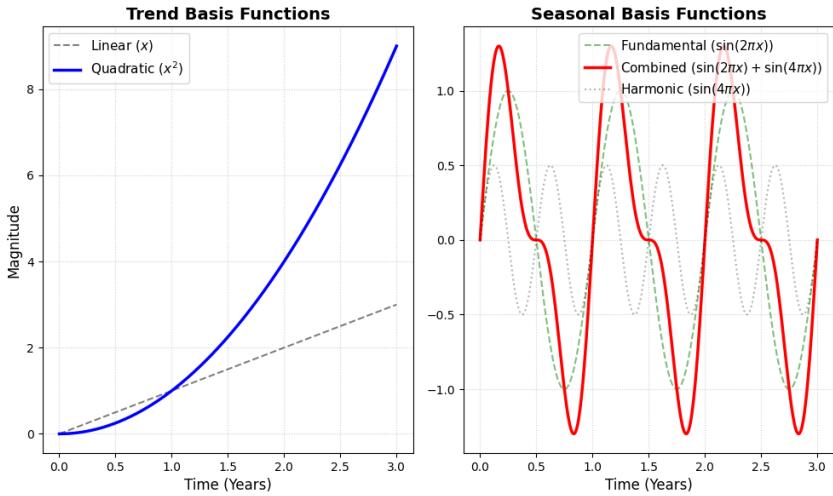


Figure 3.4: 전체적인 추세를 반영하기 위한 기저 함수(좌), CO_2 농도의 계절적 변동의 비정현파성을 반영하기 위한 기저 함수(우)

3.1.3 최종 BLR 모델($M=7$) 예측 결과 및 토의

$M=7$ 모델의 최종 예측 결과는 Figure 3.5과 같으며, 3개 모델의 수치적 성능 비교는 Table 3.1과 같다.

Table 3.1: BLR 모델별 수치적 성능 비교

모델 (기저 함수)	RMSE (ppm)	95% CI Coverage	Avg. Std. Dev (ppm)
$M=2$ (Linear)	13.2202	0.00 %	2.0074
$M=3$ (Quadratic)	3.0567	75.42 %	2.0258
$M=7$ (Quad+Seasonal)	2.0716	100.00 %	2.0315

Table 3.1은 $M=7$ 모델이 예측 정확도(Accuracy)와 불확실성 정량화(UQ) 양쪽 모두에서 매우 성공적인 결과를 달성했음을 보여준다. 각 수치적 지표의 의미는 다음과 같다.

- **예측 정확도 (RMSE):** 모델의 예측 평균(μ_*)과 실제 관측값(y_{test}) 사이의 평균 오차를

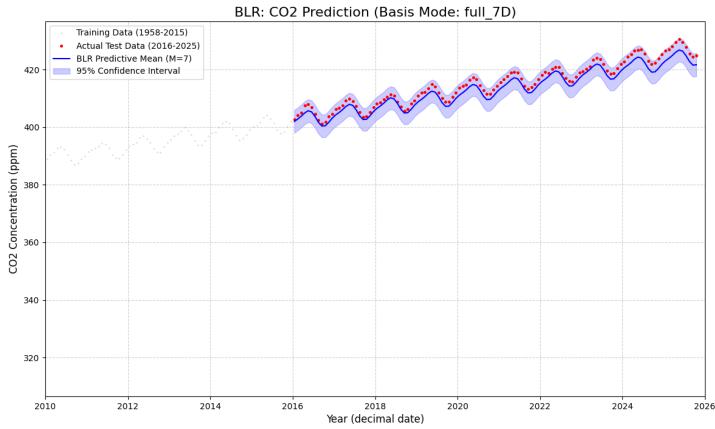


Figure 3.5: 복합 기저함수($M=7$)를 사용한 최종 BLR 예측 결과

의미하는 RMSE(Root Mean Squared Error)는 $M=7$ 모델에서 2.0716 ppm으로 세 모델 중 가장 낮게 나타났다. 이는 Figure 3.5에서 예측 평균(파란선)이 실제 데이터(빨간점)를 정확히 따라가는 것과 일치하는 결과이다.

- **불확실성 신뢰도 (95% CI Coverage):** 모델이 예측한 95% 신뢰 구간이 실제 테스트 데이터를 얼마나 포함하는지(95%가 목표치)를 나타내는 Coverage는 100.00%로, 목표치를 완벽하게 만족하였다. (100%는 95%보다 다소 보수적인 예측일 수 있으나, 0%($M=2$)나 75%($M=3$)에 비하면 매우 잘 보정된 결과이다.)
- **원인 분석 (Avg. Std. Dev):** 이러한 높은 신뢰도의 원인은, 모델이 스스로 예측한 평균 불확실성(Avg. Std. Dev: 2.0315 ppm)이 실제 평균 오차(RMSE: 2.0716 ppm)와 거의 동일하기 때문이다. 하이퍼파라미터 σ^2 가 데이터의 실제 노이즈 수준 ($\sigma^2 \approx 2.0^2 = 4.0$)에 맞게 잘 설정되었음을 의미한다.
- **결론:** $M=7$ 모델은 복잡한 기저 함수로 신호를 성공적으로 포착했으며, 적절하게 설정된 하이퍼파라미터(σ^2)를 통해 노이즈의 크기까지 정확하게 정량화하였다. 이는 BLR이 수동적이긴 하지만, (1) 기저 함수와 (2) 하이퍼파라미터가 데이터에 맞게 잘 설계된다면 매우 강력한 예측 및 UQ 성능을 보일 수 있음을 증명한다.

3.2 가우시안 프로세스(Gaussian Process, GP)

3.2.1 GP 모델링 및 커널 설계

데이터의 복잡한 패턴(추세+계절성)을 모델링하기 위해, 3가지 기본 커널을 덧셈으로 조합하여 복합 커널을 설계하였다. 이는 앞선 BLR에서 논의한 내용대로, CO₂ 데이터가 (1) 장기적인 추세, (2) 주기적인 계절성, (3) 불규칙한 노이즈의 합으로 구성된다고 가정하는 것과 동일하다.

사용된 커널의 조합은 다음과 같다.

$$k(x, x') = k_{\text{RBF}}(x, x') + k_{\text{ExpSineSquared}}(x, x') + k_{\text{White}}(x, x') \quad (3.5)$$

- **장기적인 추세(RBF Kernel):** RBF(Radial Basis Function) 커널은 CO₂ 농도의 비선형적이며 부드러운 장기적 상승 추세를 모델링한다. 이 커널은 두 시점의 시간 차이가 가까울수록 유사하고, 멀어질수록 관련성이 지수적으로 감소한다고 가정한다.
- **계절성(ExpSineSquared Kernel):** ExpSineSquared 커널은 1년 주기로 반복되는 명확한 계절성 패턴을 포착한다. 이 커널은 두 시점의 시간 차이가 1년의 배수에 가까울수록 서로 매우 유사하다고 가정한다.
- **노이즈(White Kernel):** WhiteKernel은 데이터의 불규칙한 노이즈(noise) 또는 측정 오차를 모델링한다. 이 커널은 각 데이터 포인트의 노이즈가 다른 모든 포인트와 독립적이라고 가정하며, 추세와 계절성으로 설명되지 않는 무작위적인 변동성을 흡수하는 역할을 한다.

3.2.2 커널 하이퍼파라미터 튜닝 및 모델 최적화

커널의 성능은 하이퍼파라미터 설정에 크게 의존한다. 본 보고서에서는 각 커널의 특성에 맞춰 다음과 같은 튜닝 전략을 적용하였다.

- **장기적인 추세(RBF Kernel):** `length_scale_bounds`(`length_scale`의 최적화 탐색 범위)를 (40,100)으로 설정하였다. 이는 본 모델의 가장 핵심적인 전략이다. `length_scale`(함수의 부드러움/변화 척도) 파라미터를 의도적으로 40년 이상의 매

우 큰 값으로 제한함으로써, RBF 커널이 1년 단위의 미세한 계절성 변동에 과적합 되는 것을 원천적으로 방지하였다. 그 결과, RBF 커널은 오직 수십 년 단위의 거시적이고 부드러운 추세만 학습되도록 강제되었다.

- **계절성(ExpSineSquared Kernel):** `periodicity`(함수의 반복 주기)를 1.0으로, `periodicity_bounds`(주기의 최적화 탐색 범위)를 "fixed"로 설정하였다. 이는 CO₂ 데이터의 주기가 1년이라는 지식을 모델에 반영한 것이다. 주기를 고정함으로써, 모델이 데이터의 주기를 탐색하는 데 시간을 낭비하지 않고, 더 빠르고 안정적으로 1년의 계절성 패턴을 학습하도록 유도하였다.
- **노이즈(White Kernel):** `noise_level_bounds`(노이즈 수준의 최적화 탐색 범위)를 (1e-5, 1e1)로 설정하여, 모델이 데이터의 고유한 노이즈 수준(분산)을 합리적인 범위 내에서 자유롭게 찾도록 허용하였다.

하이퍼파라미터의 최종 값은 `gp.fit()` 과정을 통해 로그-주변-가능도(Log Marginal Likelihood, LML)를 최대화하는 방향으로 자동 최적화한다. 이 최적화 과정은 Local Optimum에 빠질 위험이 있으므로, `n_restarts_optimizer = 10`으로 설정하여, 서로 다른 10개의 시작점에서 최적화를 재시도하고, 그 중 가장 높은 LML이 높은 파라미터 조합을 채택하도록 하였다.

3.2.3 실험 결과 및 분석

앞서 설계한 커널과 튜닝 전략을 기반으로 GP 모델을 학습시키고, 2016년부터 2025년까지의 테스트 데이터셋에 대한 성능을 평가하였다.

최적화된 커널 파라미터

`gp.fit()`을 통해 로그-주변-가능도(LML)를 최대화한 결과, 최종적으로 최적화된 커널의 하이퍼파라미터는 다음과 같다.

Optimized kernel:

```
RBF(length_scale=40) +
ExpSineSquared(length_scale=4.87, periodicity=1) +
WhiteKernel(noise_level=0.000548)
```

주목할 점은 RBF 커널의 `length_scale`이 40.0으로, 본 보고서에서 설정한 탐색 범위 (40, 100)의 하한값으로 최적화되었다는 것이다.

이는 RBF 커널이 장기 추세를 학습하는 과정에서 40년보다 더 짧은 `length_scale`을 선호했음을 시사한다. 하지만 (40, 100)이라는 제약조건으로 인해, 모델은 거시적 추세만 학습하도록 성공적으로 강제되었으며, 이는 의도한 튜닝 전략이 효과적으로 작동했음을 보여준다.

또한 WhiteKernel의 `noise_level`이 약 5.48×10^{-4} 라는 매우 작은 값으로 수렴한 것은, RBF와 ExpSineSquared 커널의 조합이 데이터의 변동성을 거의 완벽하게(99.9% 이상) 설명하고 있음을 의미한다.

시각적 분석

Figure 3.6은 학습된 GP 모델이 테스트 데이터셋을 예측한 결과를 시각화한 것이다.

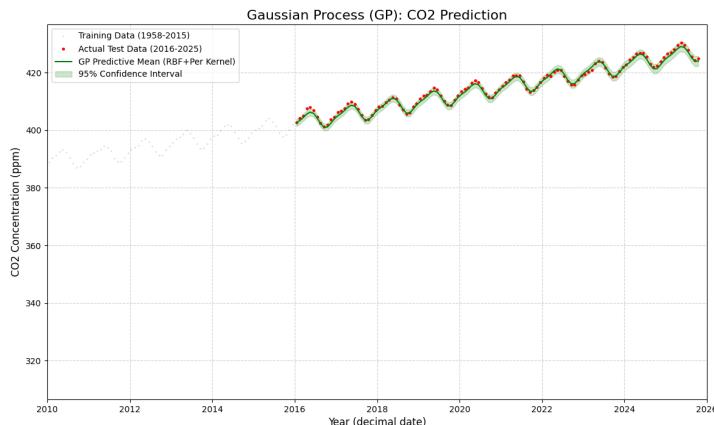


Figure 3.6: GP 모델의 CO₂ 농도 예측 결과 (2010-2026)

Figure 3.6에서 GP 예측 평균(녹색 실선)이 실제 테스트 데이터(붉은 점)의 전반적인 추세와 1년 주기의 계절성을 매우 정확하게 추종하는 것을 시각적으로 확인할 수 있다. 특히 RBF와 ExpSineSquared 커널의 조합이 장기적인 상승 추세와 주기적인 변동을 성공적으로 분해하고 모델링했음을 보여준다.

또한, 대부분의 실제 데이터(붉은 점)가 95% 신뢰구간(연녹색 음영) 내에 포함되어, 모델이 예측의 불확실성을 합리적으로 추정하고 있음을 알 수 있다.

정량적 성능 평가

모델의 예측 정확도와 불확실성 추정의 신뢰도를 정량적으로 평가하기 위해 RMSE와 95% 신뢰구간 커버리지를 측정하였다 (표 3.2).

Table 3.2: GP 모델 정량적 성능 평가

성능 지표	값
RMSE (ppm)	0.7108
95% CI Coverage (%)	96.61
Avg. Predictive Std Dev (ppm)	0.7065

테스트 데이터에 대한 RMSE는 0.7108 ppm으로 매우 낮게 나타나, 모델의 예측 정확도가 매우 높음을 입증한다.

95% 신뢰구간 커버리지는 96.61%로, 이론적인 목표치인 95%에 매우 근접한 값을 보였다. 이는 모델이 예측의 불확실성을 과소평가하거나 과대평가하지 않고, 통계적으로 매우 정확하게 추정하고 있음을 의미한다.

Chapter 4

미래 예측 및 최종 비교 분석

앞선 Chapter의 예측 성능 평가에서는 2015년까지의 데이터만을 훈련에 사용하였다. 이번 Chapter에서는 최종적인 장기 예측 결과를 도출하기 위해, 가용한 모든 데이터(1958년 ~ 2025년)를 사용하여 두 모델(BLR, GP)을 재학습시키고, 2040년까지의 CO₂ 농도를 예측하였다.

이 Chapter의 핵심 목표는 학습 데이터가 없는 장기 미래 구간(2026년 ~ 2040년)에서 BLR과 GP가 불확실성(Epistemic Uncertainty)을 얼마나 다르게 정량화하는지를 비교하는 것이다.

4.1 BLR 모델의 장기 예측 결과

Figure 4.1은 앞선 Chapter에서 도출한 M=7 기저 함수를 사용한 BLR 모델의 2040년까지의 예측 결과이다.

- **예측 평균:** 모델은 학습 기간(1958년 ~ 2025년) 동안의 추세와 계절성을 반영하여 2040년 CO₂ 농도를 약 467 ppm으로 예측하였다.
 - **불확실성 증가 패턴:** 관측 데이터가 끝나는 2025년 이후(수직 점선 오른쪽), 95% 신뢰 구간(파란 음영)은 수치적으로 미미하고 선형적인 증가세를 보인다.
수치적으로 보면(Table 4.1), 14년의 외삽 기간(2026년 ~ 2040년) 동안 표준편차(Std Dev)는 2.0164 ppm에서 2.0524 ppm으로, 총 1.79% 증가하는 데 그쳤다.
- 이러한 낮은 증가율은 BLR 모델의 근본적인 한계를 노출한다. M=7 모델은 식 3.4와

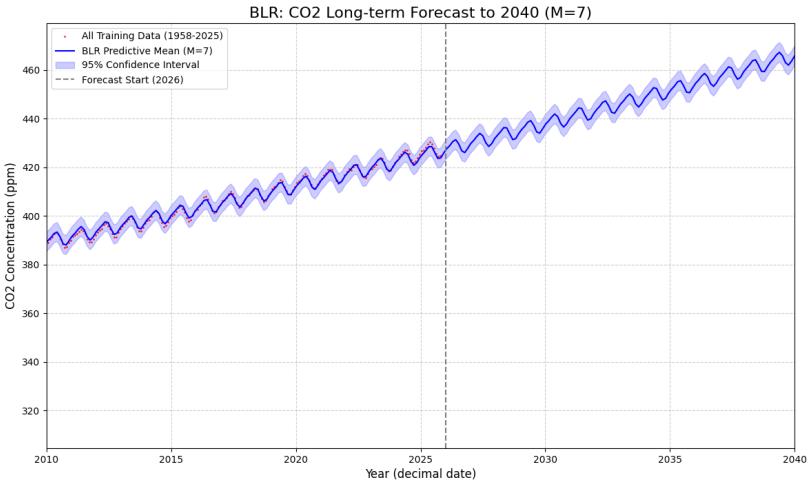


Figure 4.1: BLR M=7 모델을 이용한 CO₂ 농도 예측 (1980 ~ 2040)

같이 2차 다항식(x^2)과 유한한 주기 함수(\sin, \cos)로 구성되어 있으며, 예측 구간에서 불확실성 증가는 주기 함수의 유한성과 최고차 항인 x^2 의 영향으로 매우 완만하게 증가한다. 이는 모델이 데이터 유사성이 급격히 떨어지는 미지의 영역에서도 불확실성을 사실상 일정하게 유지하며, 예측 신뢰도를 실제보다 낮게 평가하는 결과를 낳는다.

4.2 GP 모델의 장기 예측 결과

Figure 4.2는 GP 모델의 2040년까지의 예측 결과이다.

- **예측 평균:** GP 모델은 2040년 CO₂ 농도를 약 464 ppm으로 예측하여 BLR과 유사한 중앙값을 제시하였다.
- **불확실성 증가 패턴:** 95% 신뢰구간(녹색 음영)은 2026년 이후 폭이 넓어지는 것을 시각적으로 확인할 수 있다. 수치적으로 보면(Table 4.1), GP 모델의 2026년(예측 시작 시점)의 불확실성(Std Dev)은 0.6083 ppm이었던 반면, 2040년(예측 종료 시점)의 불확실성(Std Dev)은 1.4498 ppm으로 급격히 증가했다.



Figure 4.2: GP 모델을 이용한 CO₂ 농도 예측 (1980 ~ 2040)

Table 4.1: BLR과 GP 모델의 장기 예측 불확실성 수치 비교 (2026년 ~ 2040년)

모델	Std Dev @ 2026	Std Dev @ 2040	Std Dev 증가율 (14년 간)
BLR	2.0164 ppm	2.0524 ppm	1.79 %
GP	0.6083 ppm	1.4498 ppm	138.35 %

4.2.1 최종 비교 및 토의

두 모델의 장기 예측 결과를 비교할 때, Table 4.1의 수치 분석은 아래의 결과들을 명확히 보여준다.

- **초기 정교함 비교 :** GP는 예측 시작점(2026년)에서 0.6083 ppm이라는 BLR 대비 약 3배 이상 낮은 불확실성을 보여주어, GP 모델이 커널을 통해 데이터의 신호(Signal)와 노이즈(Noise)를 더욱 정교하게 분리해냈음을 입증한다.
- **BLR의 한계 :** BLR 모델은 14년의 예측 기간 동안 1.79%의 매우 낮은 증가율을 보였다. 이는 모델이 미지의 영역에서도 불확실성이 거의 증가하지 않는다고 판단하는 것으로, 기저 함수의 구조적 한계 때문에 예측 신뢰도를 과소평가하는 결과를 낳는다.
- **GP의 장점 :** GP 모델은 동일 기간 동안 138.35%의 폭발적인 증가율을 보였다. GP는

학습 데이터(\mathbf{x})로부터의 거리가 멀어질수록 유사도(covariance)가 지수적으로 감소함을 감지한다. 이 급격한 분산 증가는 미지의 미래에 대한 모델의 무지를 정직하게 반영하는 것으로, 통계적으로 가장 합리적인 불확실성 정량화 방식임을 입증한다.

Chapter 5

결론

본 보고서는 대기 중 CO₂ 농도 시계열 데이터의 예측 및 불확실성 정량화(UQ)를 목표로, 베이즈 선형 회귀(BLR)와 가우시안 프로세스(GP)라는 두 가지 확률론적 회귀 모델을 비교 분석하였다. 두 단계의 실험을 통해 두 모델의 근본적인 차이와 장단점을 명확히 파악할 수 있었다.

1. 예측 정확도 및 초기 정교함 비교 : 첫 번째 실험에서, GP 모델은 커널 조합 및 자동 최적화 전략을 통해 RMSE 0.7108 ppm이라는 BLR 모델(RMSE 2.0716 ppm)보다 뛰어난 예측 정확도를 달성하며 초기 우위를 점했다. 또한 두 번째 실험을 통해, 예측 시작점(2026년)에서 GP의 불확실성(Std Dev 0.6083 ppm)이 BLR(Std Dev 2.0164 ppm)보다 약 3배 이상 낮게 나타나, GP가 커널을 통해 신호와 노이즈를 더 정교하게 분리해내는 초기 정교함을 입증하였다. 이는 복잡한 시계열 구조 모델링에서 GP의 유연성이 BLR의 수동적인 기저 함수 설계보다 근본적으로 우월함을 시사한다.

2. 장기 예측 불확실성 정량화 : 장기 예측 능력 비교에서는, 두 모델의 불확실성 처리 방식에서 극명한 차이가 드러났다. BLR 모델은 14년의 예측 기간 동안 불확실성이 1.79% 증가하는 데 그쳤다. 이는 모델이 기저 함수의 구조적 제약으로 인해 미지 영역에서도 불확실성이 거의 증가하지 않는다고 판단하는 과도한 자신감을 나타내며, 예측 신뢰도를 과소평가하는 한계를 노출했다. 반면, GP 모델은 동일 기간 동안 불확실성이 138.35%의 폭발적인 증가율을 보였다. GP는 학습 데이터로부터의 거리 증가에 따른 유사도 감소를 감지하여, 미지의 미래에 대한 위험도를 정직하고 보수적인 방식으로 반영하는 것을 입증하였다.

3. 환경적 의미 및 제언: 최종 예측 결과, 2040년의 CO₂ 농도는 약 464 ppm ~ 467 ppm

수준으로 예측되었다. 이 수치는 현 수준(약 420 ppm) 대비 상당한 증가를 의미하며, 국제사회가 설정한 지구 온도 상승 제한 목표를 유지하기 위한 글로벌 배출량 감축 목표를 초과하는 위험을 시사한다. 장기 예측에 있어 BLR처럼 불확실성을 과소평가하는 것은 정책 결정권자에게 잘못된 안정감을 줄 수 있다. 따라서 GP와 같이 현실적인 불확실성(예측 대역)을 제공하는 모델이 기후 변화 대응 전략의 위험 요소를 정확하게 평가하는 데 필수적이다.

결론적으로, CO₂와 같이 복잡하고 비선형적인 시계열 데이터의 예측 성능과 장기 불확실성 정량화 측면 모두에서 가우시안 프로세스 모델이 베이즈 선형 회귀 모델보다 명확하게 우위에 있음을 본 보고서에서 입증하였다.

Appendix A

전체 코드

본 보고서에서 사용된 코드들을 아래 GitHub Repository를 통해 다운받을 수 있습니다.

GitHub : https://github.com/dohjh/2025_Fall_ML/tree/main/Homework_5