APCC
APEC 기후센터

기계학습을 이용한 우리나라 여름철 장기기온예측

2021년 11월 5일
APCC 기후정보서비스 사용자워크숍

기후분석과
이진영 선임연구원

Table of CONTENTS

APCC
APEC 기후센터

- 기계학습이란
- 가우시안 프로세스 모델이란
- 우리나라 여름철 장기기온예측

기계학습이란

인공지능

인간의 지능을 필요로 하는 일들을 컴퓨터나
컴퓨터에 의해 제어되는 로봇이 수행하는 능력

Artificial intelligence

Print | Cite | Share | More

WRITTEN BY

B.J. Copeland

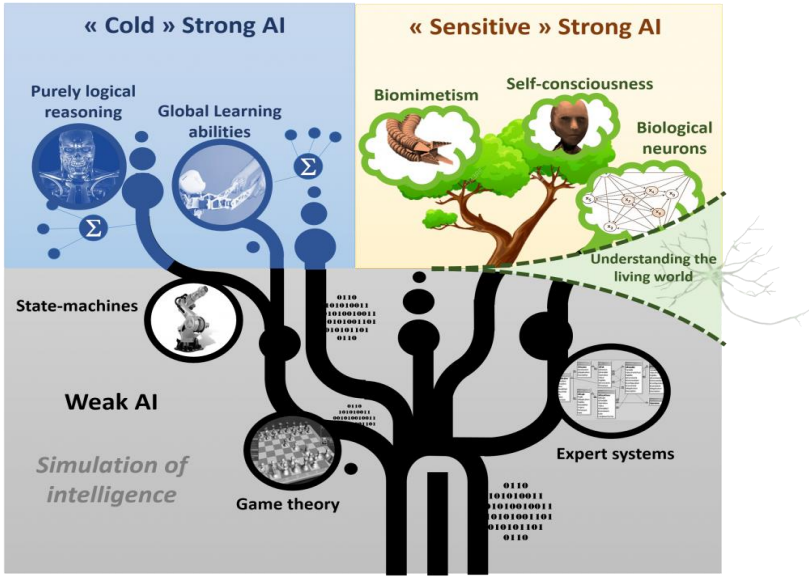
Professor of Philosophy and Director of the Turing Archive for the History of Computing, University of Canterbury, Christchurch, New Zealand. Author of *Artificial Intelligence* and others.

See Article History

Alternative Title: AI

Artificial intelligence (AI), the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience.

Strong AI는 인간을 모방하는 것 뿐 아니라 인간과
유사한 지능을 갖추고 인간처럼 사고하고 행동



이미지 출처: Théophile Gonos

Weak AI는 인간이 미리 계획한 일을 능숙히
수행하기 위한 기술을 개발하기 위해서 사고

기계학습이란

기계학습

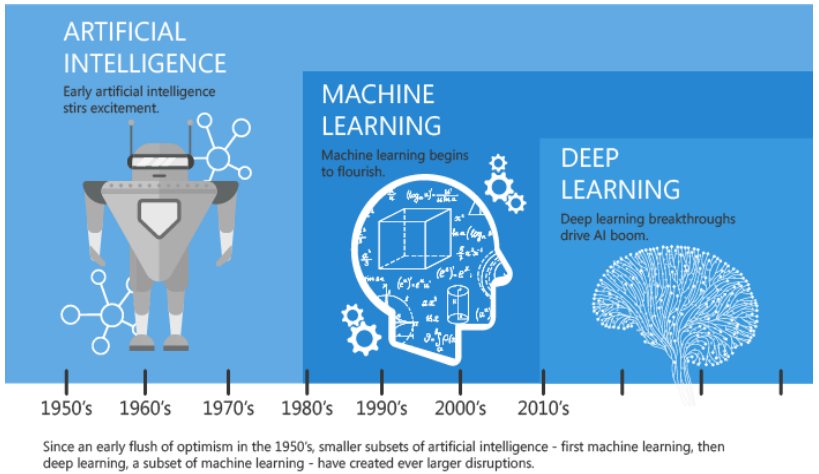
"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"

— Arthur L. Samuel, AI pioneer, 1959

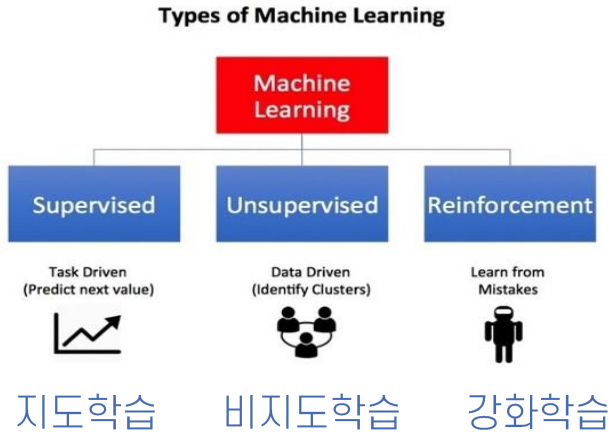
Machine learning, in artificial intelligence (a subject within computer science), discipline concerned with the implementation of computer software that can learn autonomously.

처리과정을 학습하는 컴퓨터 모델링 – 컴퓨터로 하여금 귀납이나 연역과 같은 특정 추론 전략을 사용하여 현존하는 자료나 이론으로부터 지식을 획득하게 함

출처: Jensen, 2016



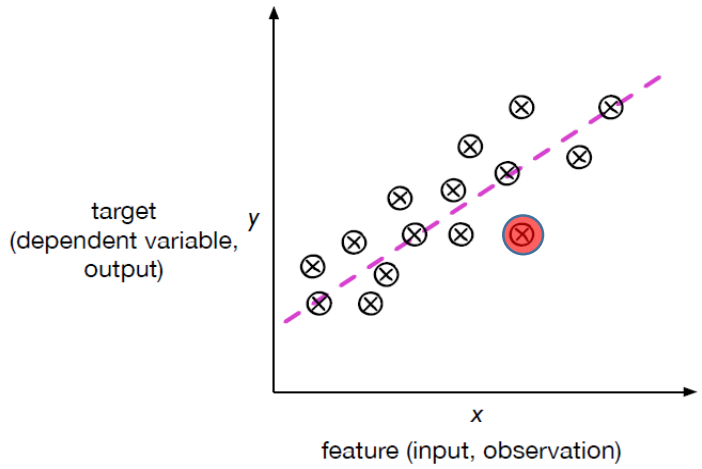
출처: NVIDIA 블로그, UNIST IRIS Lab



출처: towardsdatascience

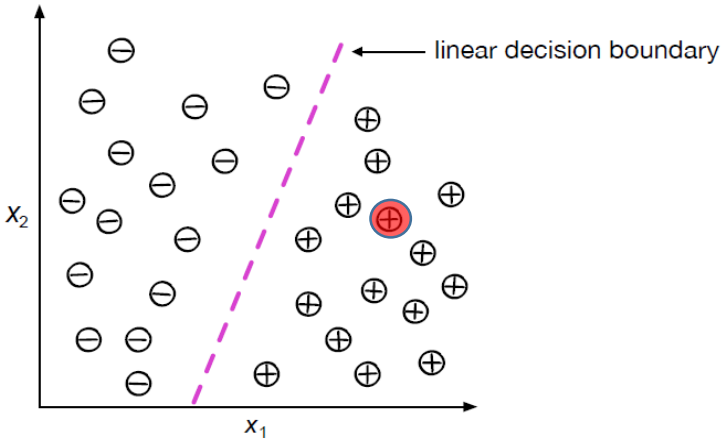
기계학습이란: 지도학습

회귀

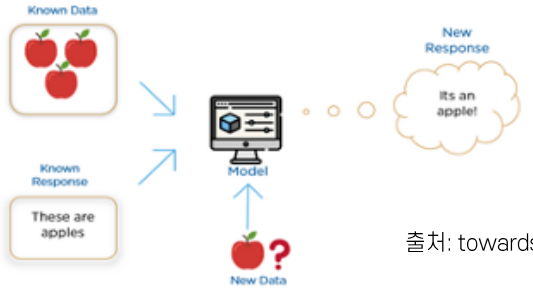


출처: Raschka and Mirjalily, 2019

분류 / 패턴인식



출처: Raschka and Mirjalily, 2019

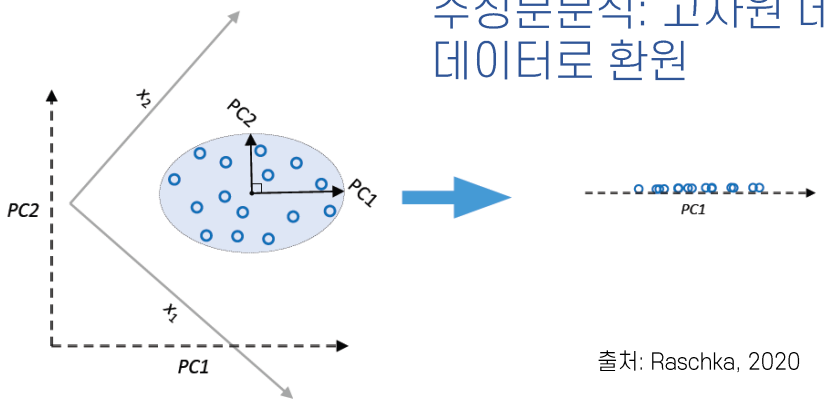


출처: towardsdatascience

기계학습이란: 비지도학습

차원 감소

주성분분석: 고차원 데이터를 저차원 데이터로 환원

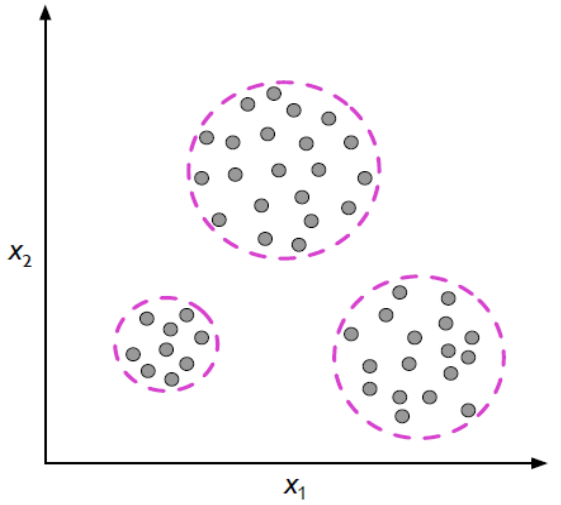


출처: Raschka, 2020

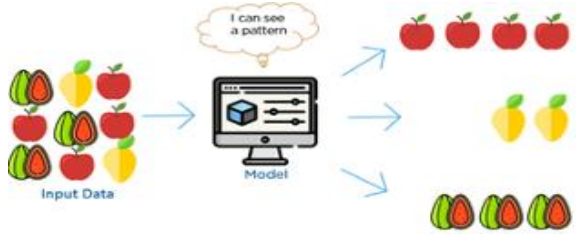
<그림 6> k개의 eigenface만을 이용한 데이터 복원(reconstruction)

출처: darkpgmr 블로그

군집

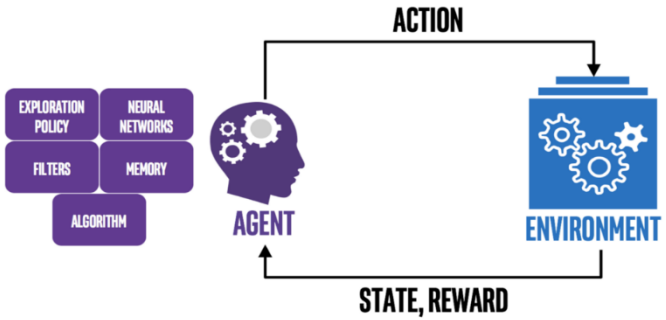


출처: Raschka, 2020



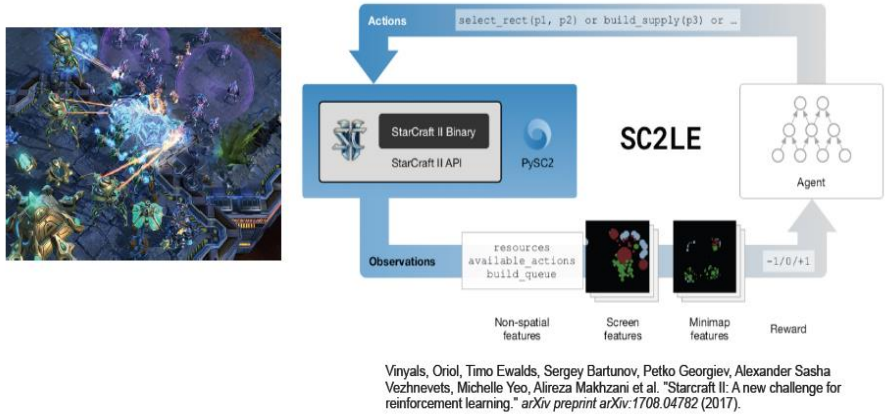
출처: towardsdatascience

기계학습이란: 강화학습



출처: medium.com

Starcraft II



Vinyals, Oriol, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani et al. "Starcraft II: A new challenge for reinforcement learning." *arXiv preprint arXiv:1708.04782* (2017).

어떤 환경 안에서 정의된 에이전트(행위자)가 현재의 상태를 인식하여 선택 가능한 행동들 중 보상을 최대화하는 행동을 선택하는 방법

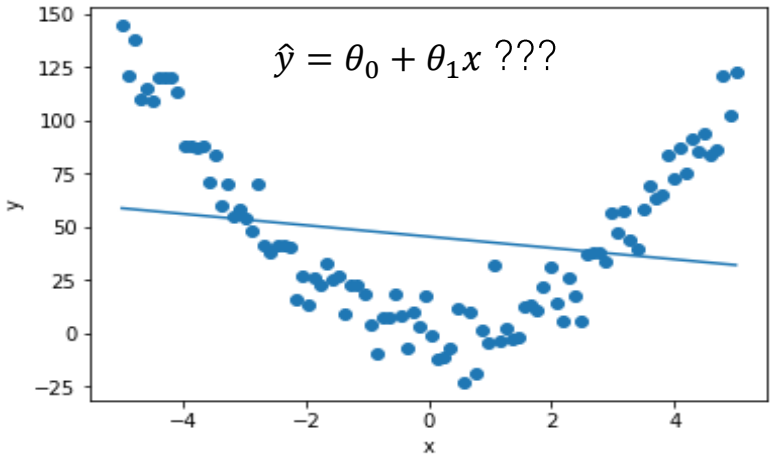
기계학습이란

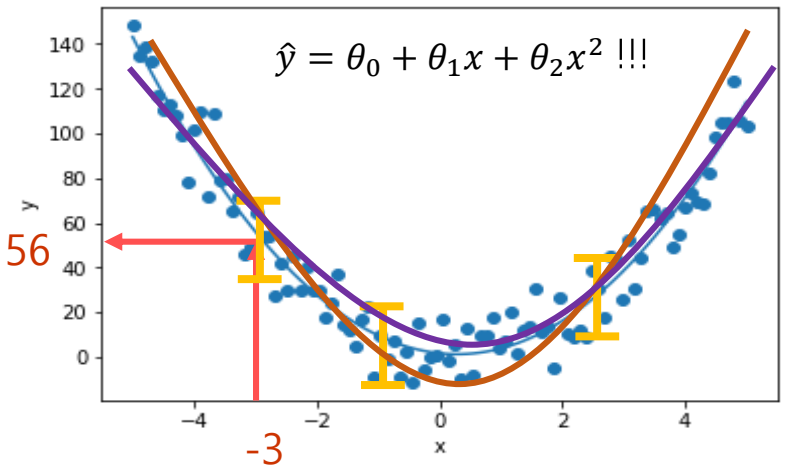

지도학습 활용의 단계
풀고 싶은 문제를 정의하자
자료를 수집하자
활용할 기법을 선정하자
모델 학습을 위한 최적화 방법을 정하자
모델 평가를 위한 기준을 정하자

가우시안 프로세스 모델이란

가우시안 프로세스: 비모수 베이지안 모델

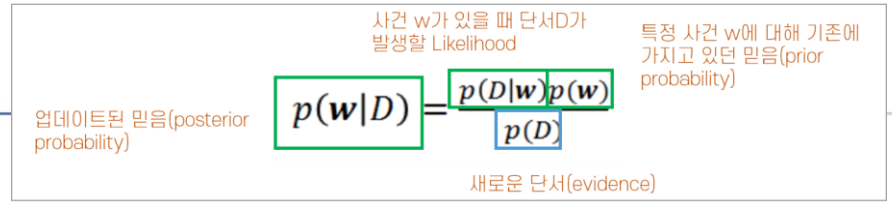
전통적 추론(Classical Inference)





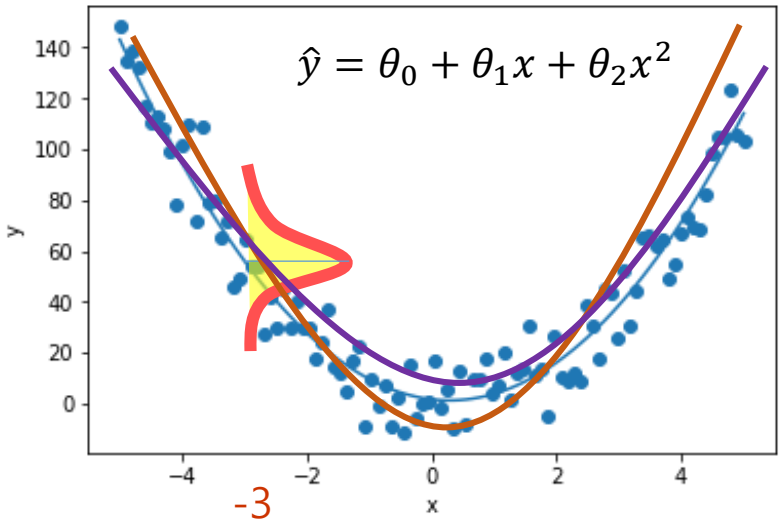
- 우리의 목적: 새로운 X_*에 대한 y 값 추정하도록 → 모집단의 특성을 파악(즉 적합한 함수 도출)
- 전통적 추론: 관측된 자료(훈련 자료) 에 관한 함수를 사용하여 그 함수가 가지는 미지의 모수를 추론
- 하지만 전통적 추론에 의한 결과는 자료에 대한 불확실성 정보를 알려주지 않음!

가우시안 프로세스 모델이란



가우시안 프로세스: 비모수 베이지안 모델

베이지안 추론(Bayesian Inference)



- 우리가 추론하고자 하는 모수는 불확실하며 이 불확실성의 정도를 확률모델로 표현하고 싶음
- 모수 각각에 대한 확률모델과 관측된 자료를 사용
- 확률예측을 통해 불확실성 정보를 제공

관측된 자료 y 에 대한 확률모델

사후확률 $p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{p(y)}$ 사전확률

- 모수에 관한 과거의 경험이나 사전 지식 같은 주관적 견해를 수량화한 모수의 특성(사전확률분포)을
- 자료로부터 얻은 모수에 관한 정보(관측된 자료)와 결합하여
- 사후확률분포를 얻음(확률예측결과를 제공)

$\theta_0, \theta_1, \theta_2$ 각각의 사전확률분포
입력 X 대해 관측된 자료 y

↓

$\theta_0, \theta_1, \theta_2$ 각각의 사후확률분포

↓

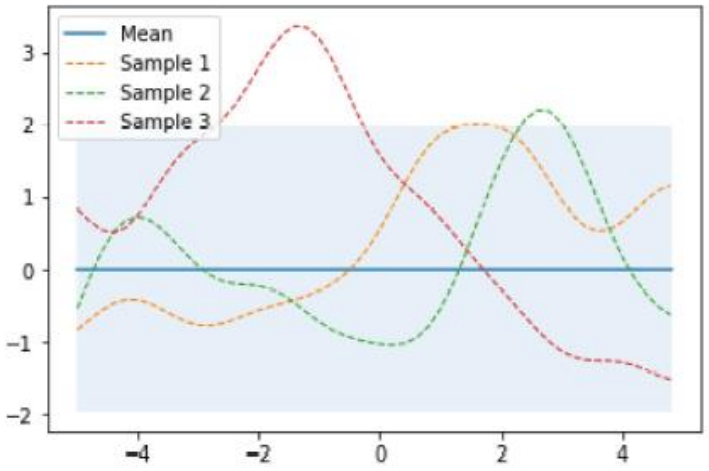
새로운 X_*에 대한 예측 값 $\hat{y}$의 분포

가우시안 프로세스 모델이란

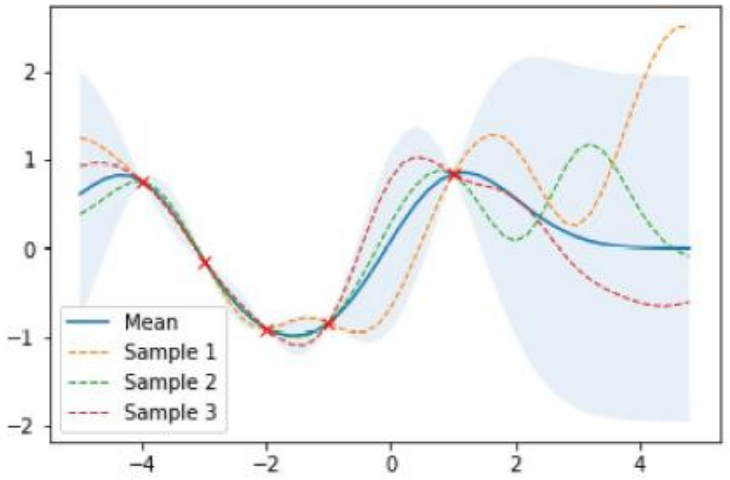
가우시안 프로세스: 비모수 베이지안 모델

- 앞서 본 예시처럼 어떤 함수를 사용할지(모수가 몇 개인지, 모수가 어떤 분포를 따를지) 규정하는 대신
- "모든 가능한 함수"에 사전확률을 주고 더 그럴듯한 함수에 높은 사후확률을 주면 어떨까?

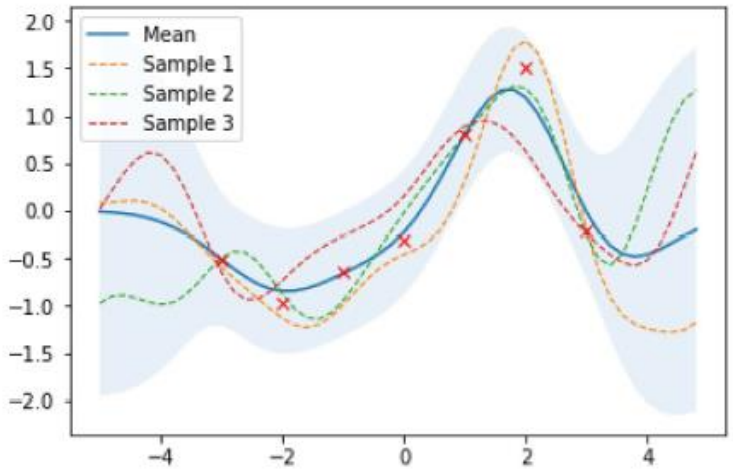
- 각각의 함수를 특정한 입력 x 에서 함수값 f(x)를 가지는 굉장히 긴 벡터로 생각하고,
- 무한한 모든 함수를 고려할 필요 없이 관측된 유한한 지점에서의 함수값의 특성만을 보면 됨



Three samples from prior



Three samples from posterior
Noise-free data

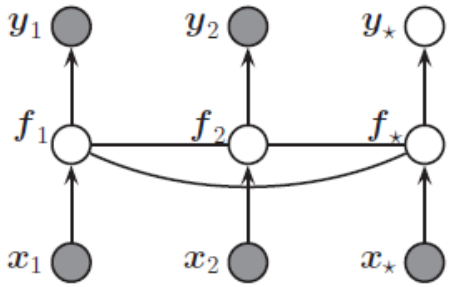


Three samples from posterior
Noisy data

출처: martin Krasser github

가우시안 프로세스 모델이란

Gaussian Process는 확률과정(stochastic process)으로 임의의 점 $x \in R^d$ 에서 확률변수 $f(x)$를 가지며 유한한 확률변수들의 결합확률분포 $p(f(x_1), \ldots, f(x_N))$ 는 Gaussian



$$p(f|X) = N(f|\mu, \mathrm{K})$$

$\mu(x)$ 평균 함수

$K_{ij} = \kappa(x_i, x_j)$ 공분산함수(커널)

x_i 와 x_j 가 가까우면 $f(x_i)$와 $f(x_j)$도 가까울 것이라 가정 ➡ 공분산함수(커널)에 의해 결정

가우시안 프로세스 모델이란

이미 관측한 자료 y와 예측 f_* 의 결합분포

$$\begin{pmatrix} y \\ f_* \end{pmatrix} \sim N\left(0, \begin{pmatrix} K_y & K_* \\ K_*^T & K_{**} \end{pmatrix}\right)$$

$$K_y = \kappa(X, X) = K$$

$$K_* = \kappa(X, X_*)$$

$$K_{**} = \kappa(X_*, X_*)$$

새로운 X_*에 대한 예측 f_*

$$\mathrm{p}(f_*|\mathrm{X}_*, X, y) = \int p(f_*|f, \mathrm{X}_*)p(f|X, y)df$$

$$\mathrm{p}(f_*|\mathrm{X}_*, X, y) = N(f_*|\mu_*, \Sigma_*)$$

$$\mu_* = K_*^T K_y^{-1} y$$

$$\Sigma_* = K_{**} - K_*^T K_y^{-1} K_*$$

RBF 커널 및 매개변수

$$\kappa(x_i, x_j) = \sigma_f^2 \exp\left(-\frac{1}{2}(x_i - x_j)^T M(x_i - x_j)\right) + \sigma_y^2 \delta_{ij}$$

Noise variance

Vertical scale parameter

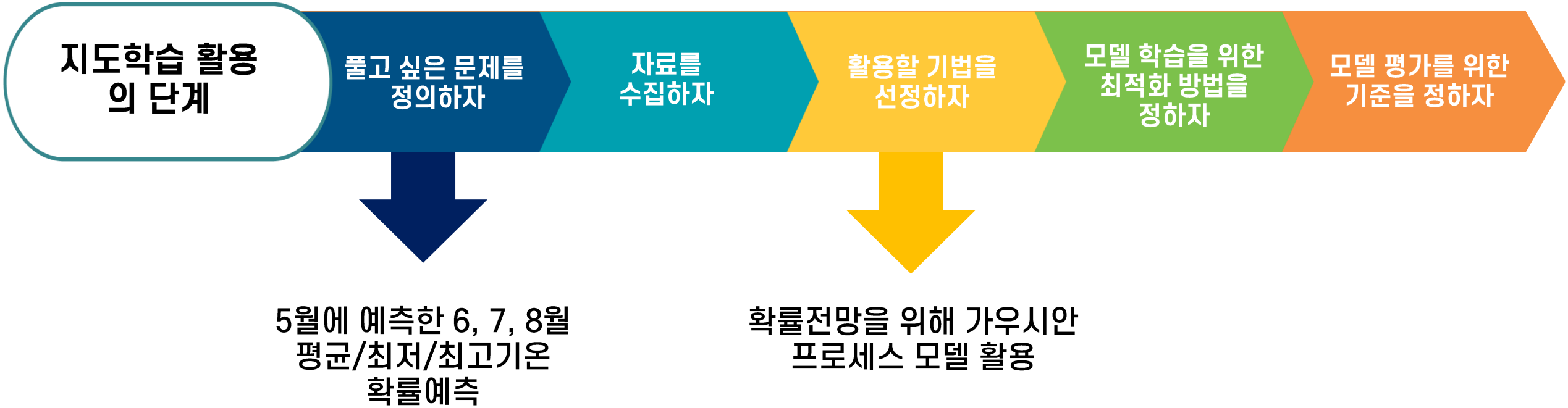
$$M = l^{-2} I \quad or \; M = diag(l)^{-2}$$

Length scale parameter

로그우도함수 최대화

$$\log p(y|X) = -\frac{1}{2} y^T K_y^{-1} y - \frac{1}{2} \log|K_y| - \frac{N}{2} \log(2\pi)$$

우리나라 여름철 장기기온예측



우리나라 여름철 장기기온예측

입력자료 X

관측자료 y

- 훈련기간 + 검증기간의 6, 7, 8월 평균/최저/최고기온

→ 전국 62개 ASOS 지점 자료 평균으로 산정

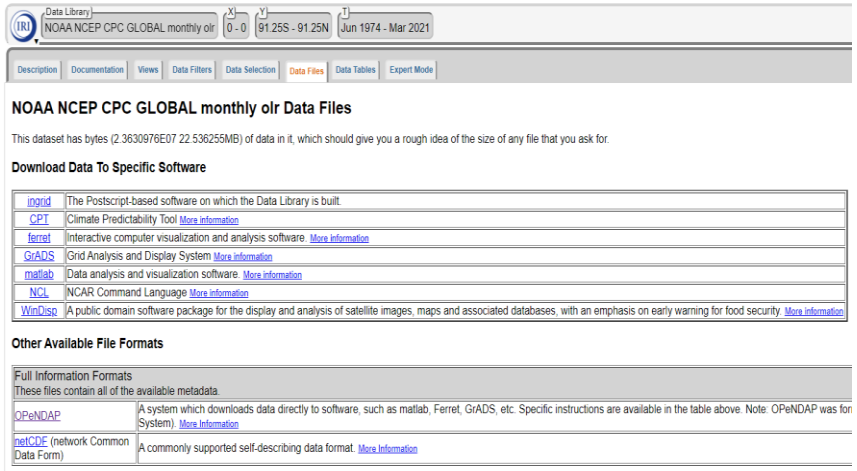
우리나라 여름철 장기기온예측

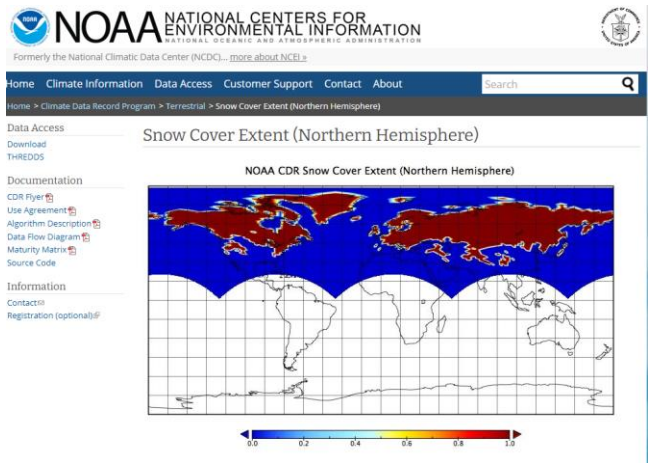
입력자료 X

관측자료 y

- Outgoing Long wave Radiation (OLR)
- Sea Surface Temperature (SST)
- Precipitation (PRCP)
- Snow Cover Extent (SCE)
- 500hPa & 850 hPa Geopotential Height (Z500, Z850)
- Sea Ice Area (SIA)
- NOAA Teleconnection, Atmospheric, and SST Indices
- MME SST, PRCP, and Z500 for periods with no obs. data

우리나라 여름철 장기기온예측





Home » ERSST_V3n

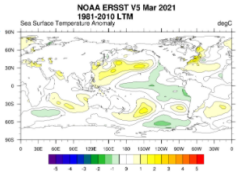
On this page: Temporal Coverage | Spatial Coverage | Levels | Update Schedule | Download/Plot Data | Analysis Tools
Restrictions | Details | Caveats | File Naming | Citation | References | Original Source | Contact

NOAA Extended Reconstructed Sea Surface Temperature (SST) V5

Values from 2008-2018 have chaned at the source. We are changing our data updates to update older values instead of just appending values each month (as of 2020/03/23).

Brief Description:

- A global monthly SST analysis from 1854 to the present derived from ICOADS data with missing data filled in by statistical methods. *More Details...*



Temporal Coverage:

- Monthly values for 1854/01 - present.
- Long term monthly means, derived from data for years 1981 - 2010.

Spatial Coverage:

- 2.0 degree latitude x 2.0 degree longitude global grid (89x180).
- 88.0N - 88.0S, 0.0E - 358.0E.

Levels:

- Sea Surface

Update Schedule:

- Variable

Download/Plot Data: (Download Issues)

Variable	Statistic	Level	Units	Download File	Create Plot/Subset
Sea Surface Temperature	Monthly Mean	Surface	degC	sst.mnmean.nc	
Sea Surface Temperature	Monthly Long Term Mean	Surface	degC	sst.mon.ltm.1981-2010.nc	

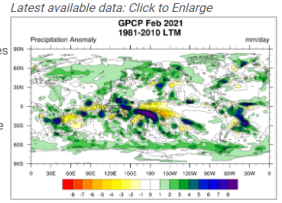
Home » PSL Home » GPCP V2 Precipitation

On this page: Temporal Coverage | Spatial Coverage | Levels | Update Schedule | Download/Plot Data | Analysis Tools
Restrictions | Details | Caveats | File Naming | Citation | References | Original Source | Contact

GPCP Version 2.3 Combined Precipitation Data Set

Note: This dataset has been updated to version 2.3 and will be updated regularly. 10/09/2020: Grids from years since 2015 have been replaced. See the NCEI webpage for more information

Latest available data: Click to Enlarge



Brief Description:

- Global Precipitation Climatology Project monthly precipitation dataset from 1979-present combines observations and satellite precipitation data into 2.5°x2.5° global grids.

Temporal Coverage:

- Monthly values 1979/01 through Feb 2021 (some months are interim).
- Long term monthly means, derived from years 1981 - 2010.

Spatial Coverage:

- 2.5 degree latitude x 2.5 degree longitude global grid (144x72)
- 88.75N - 88.75S, 1.25E - 358.75E

Levels:

- N/A

Update Schedule:

- Monthly

Home » Gridded Climate Data » NCEP Reanalysis Products: Pressure level variables

On this page: Temporal Coverage | Spatial Coverage | Levels | Update Schedule | Download/Plot Data | Analysis Tools
Restrictions | Details | Caveats | File Naming | Citation | References | Original Source | Contact

NCEP/NCAR Reanalysis 1: Pressure

We have transitioned the data files from netCDF3 to netCDF4-classic format on Monday Oct 20th, 2014.

Other Grid Types: **NCEP Reanalysis Main Page** | **Pressure Level Data** | **Surface Data** | **Surface Flux Data** | **Other Flux Data** |**Spectral Coefficents Data** | **Tropopause Data**

Brief Description:

- NCEP/NCAR Reanalysis 1

Temporal Coverage:

- 4-times daily, daily and monthly values for 1948/01/01 to present
- Long term monthly means, derived from data for years 1981 - 2010
- Values are instantaneous at the time indicated in the files

Spatial Coverage:

- 2.5 degree x 2.5 degree global grids (144x73)
- 0.0E to 357.5E, 90.0N to 90.0S

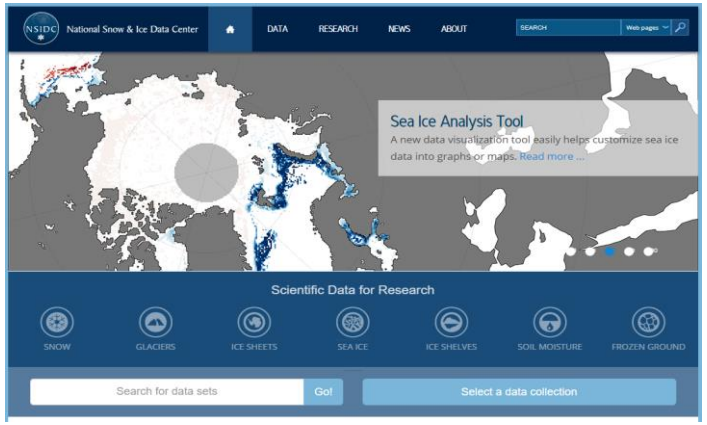
Levels:

- 17 Pressure levels (mb): 1000,925,850,700,600,500,400,300,250,200,150,100,70,50,30,20,10
- Some variables have less: omega (to 100mb) and Humidities (to 300mb).

Update Schedule:

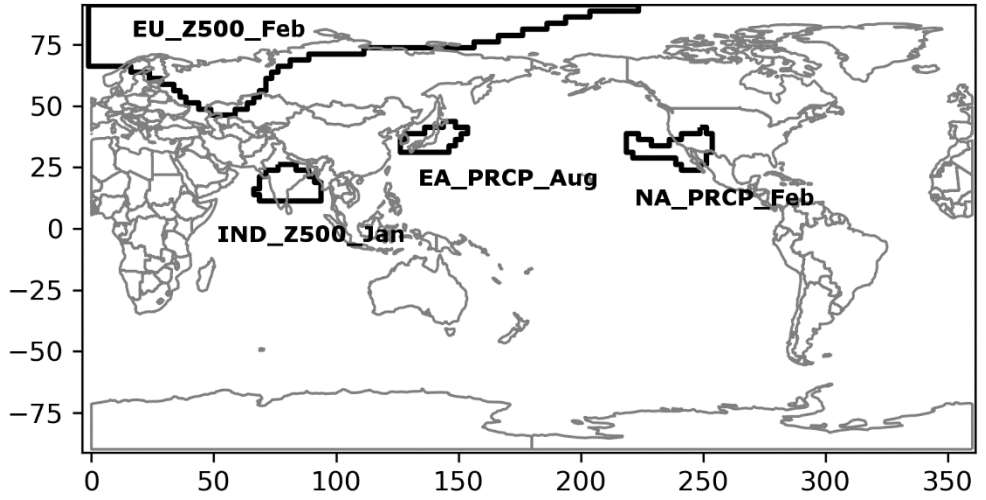
- Daily

Download/Plot Data: (Download Issues)

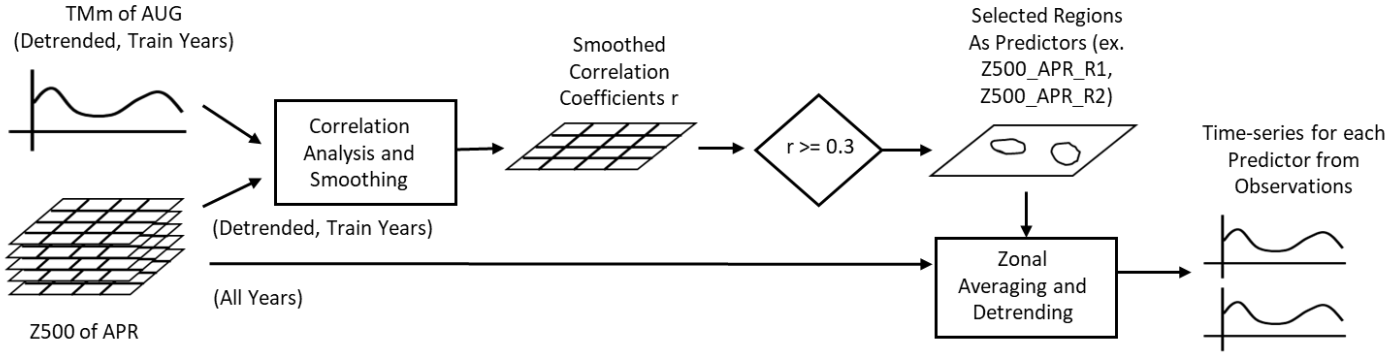


우리나라 여름철 장기기온예측

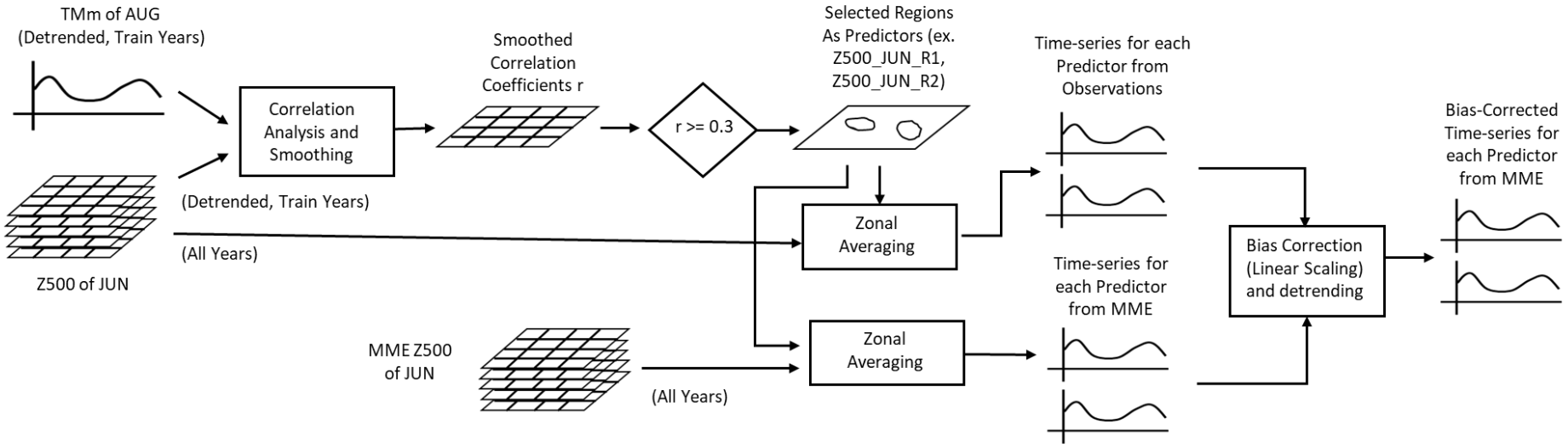
예측인자의 객관적 도출



Selection of predictors from observations: Example for target variable TMm (target month: AUG) and predictor variable Z500 (monitoring month: APR)



Selection of predictors from MME: Example for target variable TMm (target month: AUG) and predictor variable Z500 (monitoring month: JUN)



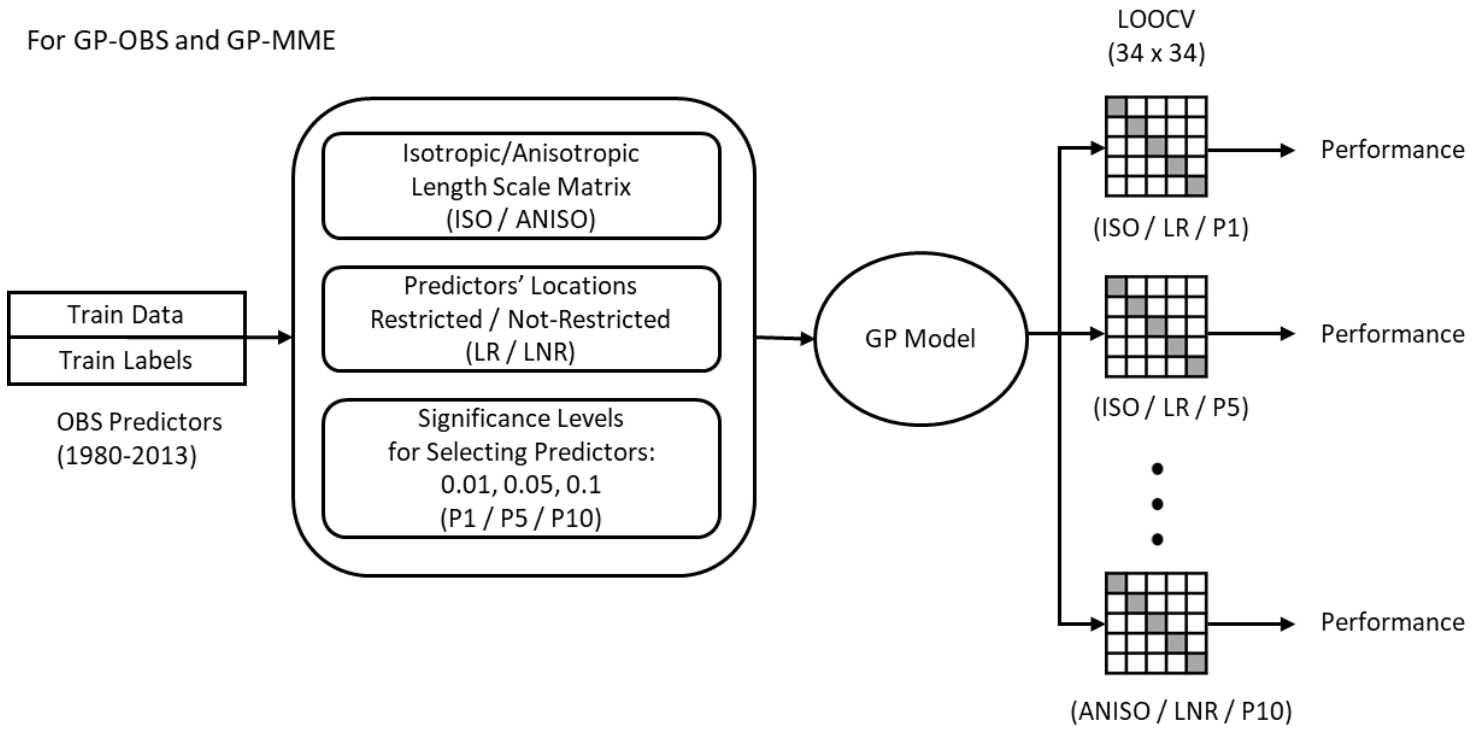
출처: Rhee and Myoung, 2021

우리나라 여름철 장기기온예측

지도학습 활용의 단계

1. 풀고 싶은 문제를 정의하자
2. 자료를 수집하자
3. 활용할 기법을 선정하자
4. 모델 학습을 위한 최적화 방법을 정하자
5. 모델 평가를 위한 기준을 정하자

초매개변수/조건 선정



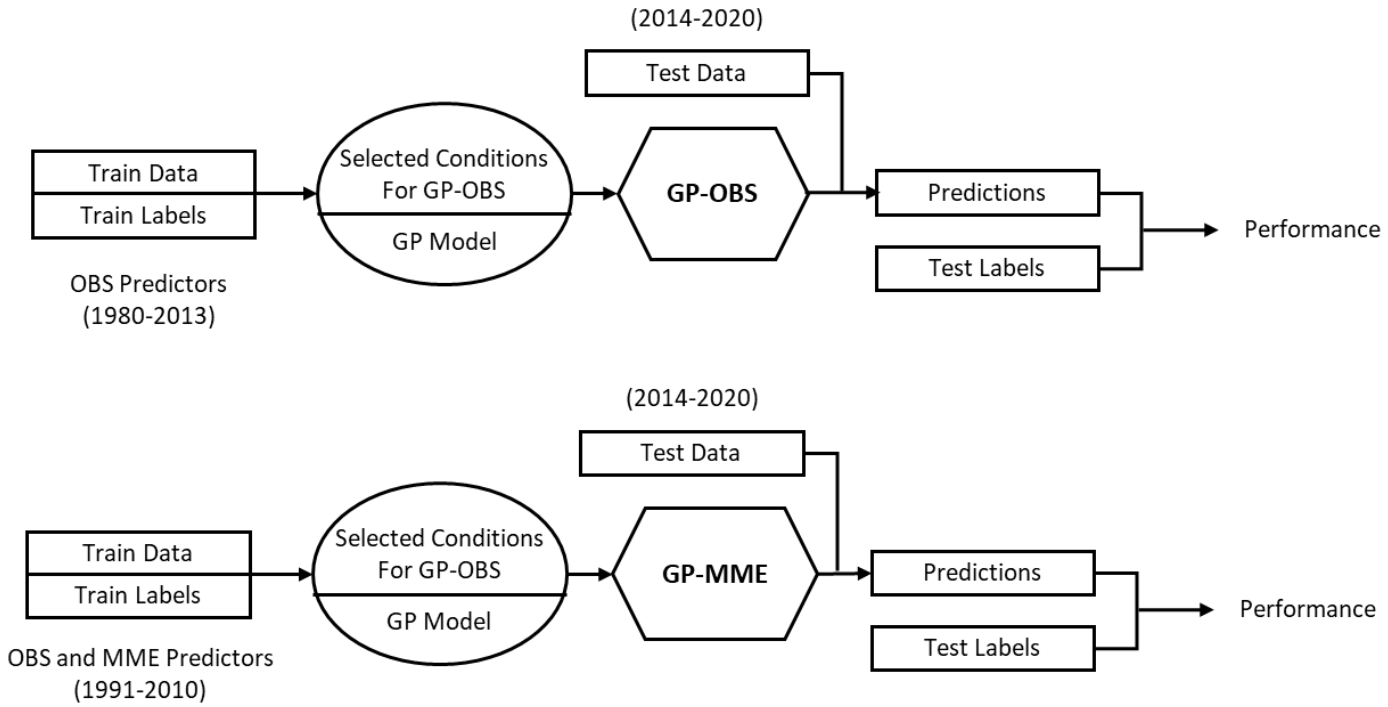
출처: Rhee and Myoung, 2021

우리나라 여름철 장기기온예측

지도학습 활용의 단계

1. 풀고 싶은 문제를 정의하자
2. 자료를 수집하자
3. 활용할 기법을 선정하자
4. 모델 학습을 위한 최적화 방법을 정하자
5. 모델 평가를 위한 기준을 정하자

모델 평가 방법 및 기준



출처: Rhee and Myoung, 2021

우리나라 여름철 장기기온예측

모델 평가 기준

Predicted Category	Observed Category			
	AN	NN	BN	
AN	a	b	c	$PRED_{AN}$
NN	d	e	f	$PRED_{NN}$
BN	g	h	i	$PRED_{BN}$
	OBS_{AN}	OBS_{NN}	OBS_{BN}	Total, n

Proportion Correct (PC)

$$\mathrm{PC} = \sum_k p(PRED_k, OBS_k) = \frac{a+e+i}{n}$$

- = Total Accuracy
- 맞게 분류된 자료 수 / 전체 자료 수
- Perfect Forecast: 1.0
- Random Forecast: 0.33

Heidke Skill Score (HSS)

$$\mathrm{HSS} = \frac{\sum_k p(PRED_k, OBS_k) - \sum_k p(PRED_k)p(OBS_k)}{1 - \sum_k p(PRED_k)p(OBS_k)}$$

- 우연히 맞게 분류되었을 경우 대비 맞게 분류된 자료 수
- Perfect Forecast: 1.0
- Random Forecast: 0.0

$$\sum_k p(PRED_k)p(OBS_k) = \left(\frac{a+b+c}{n}\cdot\frac{a+d+g}{n}\right) + \left(\frac{d+e+f}{n}\cdot\frac{b+e+h}{n}\right) + \left(\frac{g+h+i}{n}\cdot\frac{c+f+i}{n}\right)$$

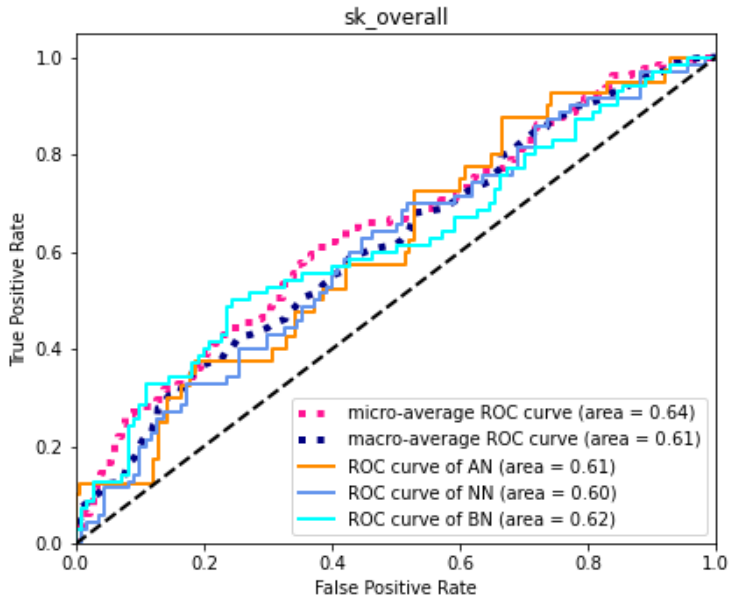
$\sum_k p(PRED_k, OBS_k) = PC$ $\qquad$ $\sum_k p(PRED_k)p(OBS_k)$ *is the random proportion correct*

우리나라 여름철 장기기온예측

모델 평가 기준

Area Under the Receiver Operating Characteristic Curve (AUC)

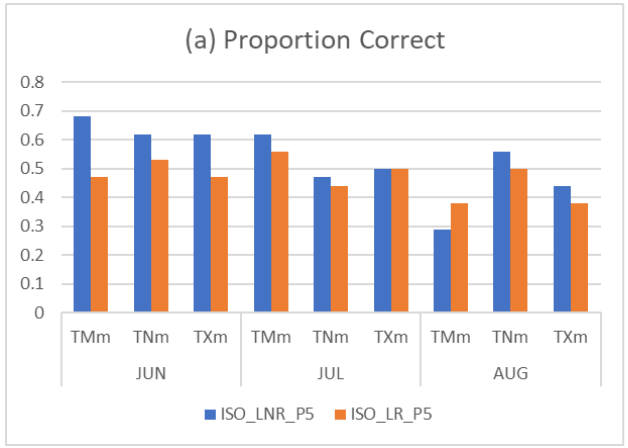
- False positive rate을 x축에, True positive rate을 y축에 둔 ROC 곡선 아래 면적
- Perfect Forecast: 1.0
- Random Forecast: 0.5

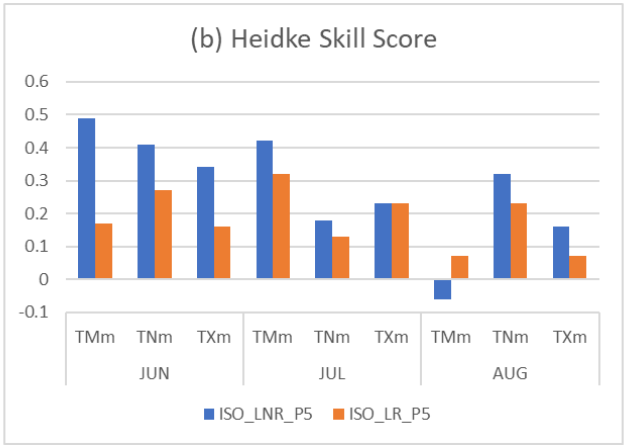


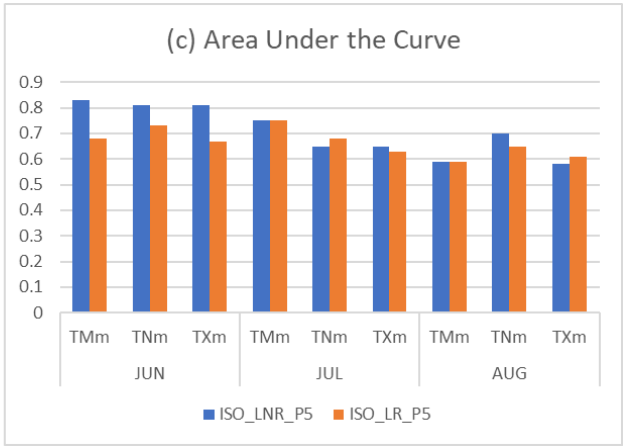
우리나라 여름철 장기기온예측

모델 평가 결과

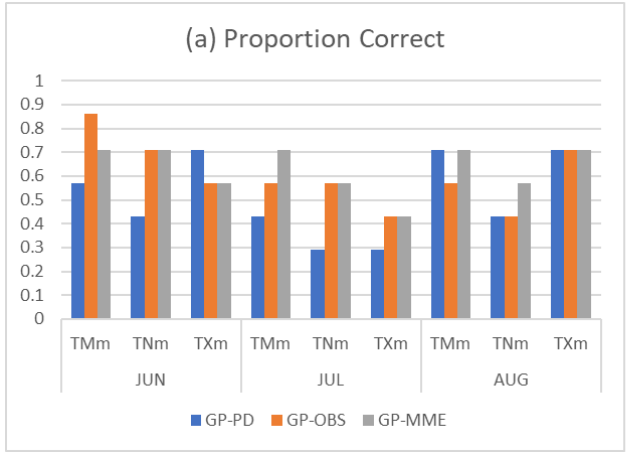
모델 조건 선정을 위한
LOOCV (1980-2013)

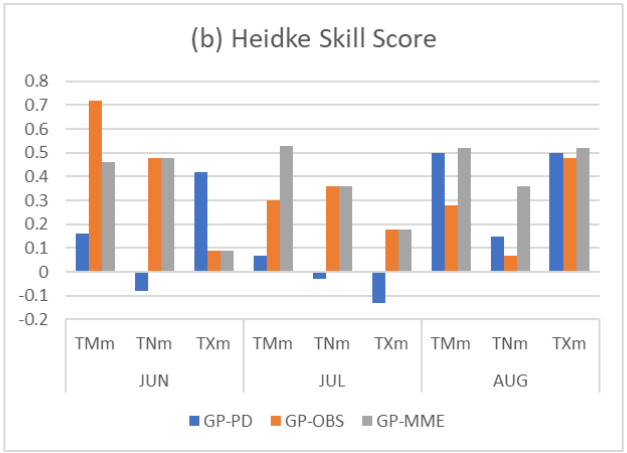


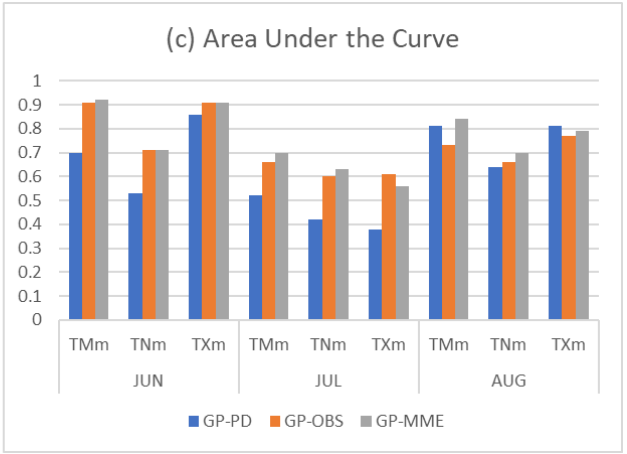




모델 검증 (2014-2020)



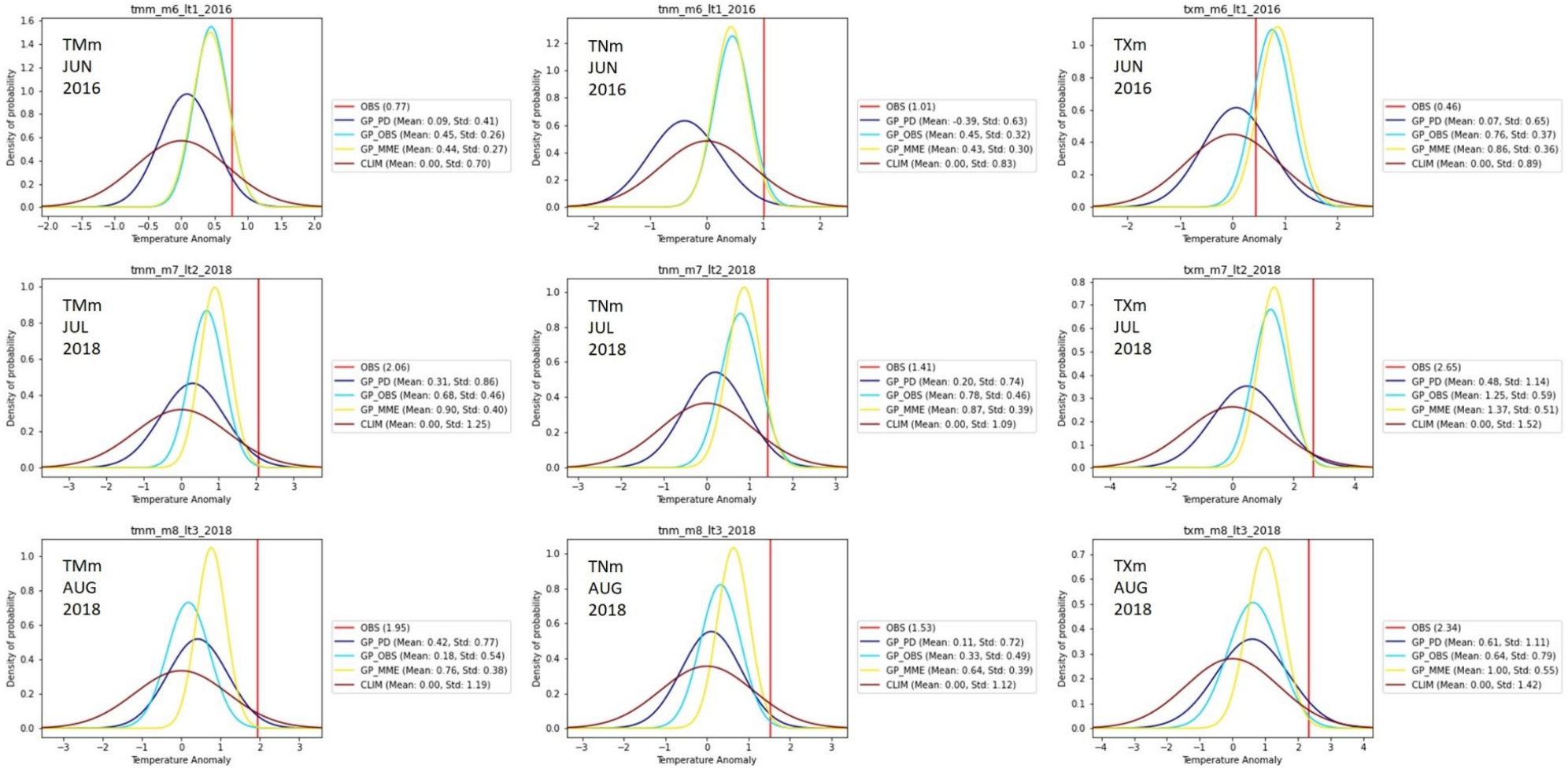




출처: Rhee and Myoung, 2021

우리나라 여름철 장기기온예측

모델 평가 결과: 개선 사례



출처: Rhee and Myoung, 2021


APCC
APEC 기후센터
감사합니다.