제4회 해양과학 빅데이터 경진대회 해양기후예측센터(인도양) 문제

SARIMA와 XGBOOST를 활용한 SCTR 지역 수온약층 깊이 지수 예측

폭풍대기: 신지은, 오승옥

[이 논문은 2022년 해양수산부 재원으로 한국해양과학기술진흥원의 지원을 받아 수행된 연구임(인도양 한-미 공동관측 및 연구, RS-2022-KS221662)]

TABLE OF CONTENTS

1 Introduction

● 연구 배경 및 목표

- **Seychelles-Chagos Thermocline Ridge(SCTR)**
  - ➢ 5–15°S 와 50–80°E에 위치한 열대 남인도양의 용승 지역으로, 수온약층이 평균 30미터로 얕다
  - ➢ SCTR 지역의 용승은 높은 영양 염류를 공급하여 주변 국가들의 주요 어장을 형성한다
  - ➢ SCTR 지역은 인도양 몬순과 매든-줄리안 진동(MJO), 엘니뇨 남방진동(ENSO) 등 기후변동성과 연관된다
  - ➢ SCTR 수온약층 변화는 해수면 온도(SST) 변동을 조절하며, 이는 대기 순환과 인도양 기후 변동성을 결정한다

- SCTR 지역 수온약층 깊이는 **(1) 바람응력 패턴과 (2) 아주 먼 거리의 기후변동성**에 의해 결정된다
  - ➢ SCTR 지역 수온약층 깊이 예측은 **기후 예측 능력 향상, 해양 생태계 이해, 해양 자원 관리 등** 다양한 분야에서 중요한 의미를 가지며, **인도양 기후 변동성을 체계적으로 이해**하는 데 기여할 수 있다

- **통계모형과 머신러닝 모델을 활용하여 2024년 5월부터 2025년 4월까지 SCTR 지역 수온약층 깊이 지수를 예측하고자 한다**

2 EDA

● 데이터 세트

지정데이터 (199101-202404)
D20_SCTR

ERA5 monthly averaged data (199101-202408)	
10m wind speed	Surface net solar radiation
10m u-component of wind	Surface net thermal radiation
10m v-component of wind	Surface latent heat flux
Sea surface temperature	Surface sensible heat flux
Mean total precipitation rate	

NOAA Monthly Climate/Ocean Indices	
Multivariate ENSO Index, MEI V2 (199101-202408)	Arctic Oscillation, AO (199101-202408)
Dipole Mode Index, DMI (199101-202404)	North Atlantic Oscillation, NAO (199101-202401)
Southern Oscillation Index, SOI (199101-202402)	Pacific Decadal Oscillation, PDO (199101-202408)

ERA5 파생변수 (199101-202408)	
U current	V current
Total heat flux	

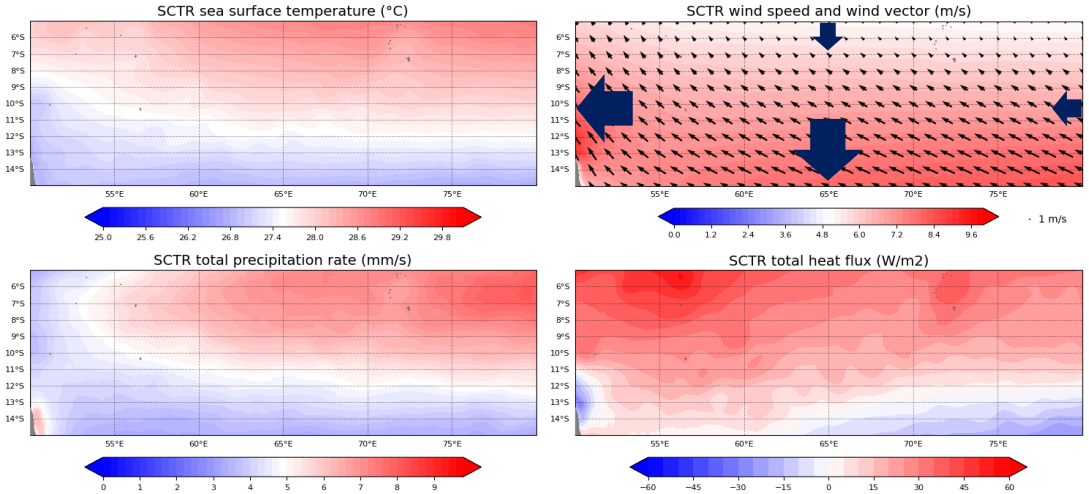
2 EDA

● 대기 및 해양 변수

$$\underbrace{\frac{\partial T_a}{\partial t}}_{\text{Tendency}} = \underbrace{-\left(u_a\frac{\partial T_a}{\partial x}+v_a\frac{\partial T_a}{\partial y}\right)}_{\text{Horizontal advection}} + \underbrace{\kappa_H\left(\frac{\partial^2 T_a}{\partial x^2}+\frac{\partial^2 T_a}{\partial y^2}\right)}_{\text{Horizontal mixing}} - \underbrace{\frac{1}{h}\left[\kappa_Z\frac{\partial T}{\partial z}\right]_{-h}}_{\text{Vertical mixing}}$$

$$\underbrace{-\left(\frac{T_a-T_{-h}}{h}\right)\left(\underbrace{\frac{\partial h}{\partial t}}_{\text{ML tendency}}+\underbrace{u_{-h}\frac{\partial h}{\partial x}+v_{-h}\frac{\partial h}{\partial y}}_{\text{Lateral induction}}+\underbrace{w_{-h}}_{\text{Vertical advection}}\right)}_{\text{Entrainment}}+\underbrace{\frac{q_0-q_{pen}}{\rho_0 c_p h}}_{\text{Net heat flux}}.$$

Vijith, V. et al. (2020)



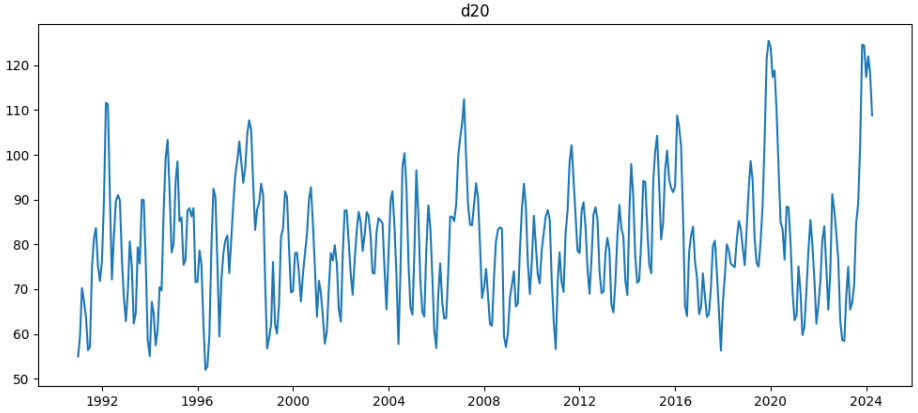
- 수온약층 깊이는 SST와 수평 이류와 수평 및 연직 혼합, Net heat flux와 관련되며, 이를 **SST, Wind, Precipitation, Heat flux**를 통해 반영하고자 하였다

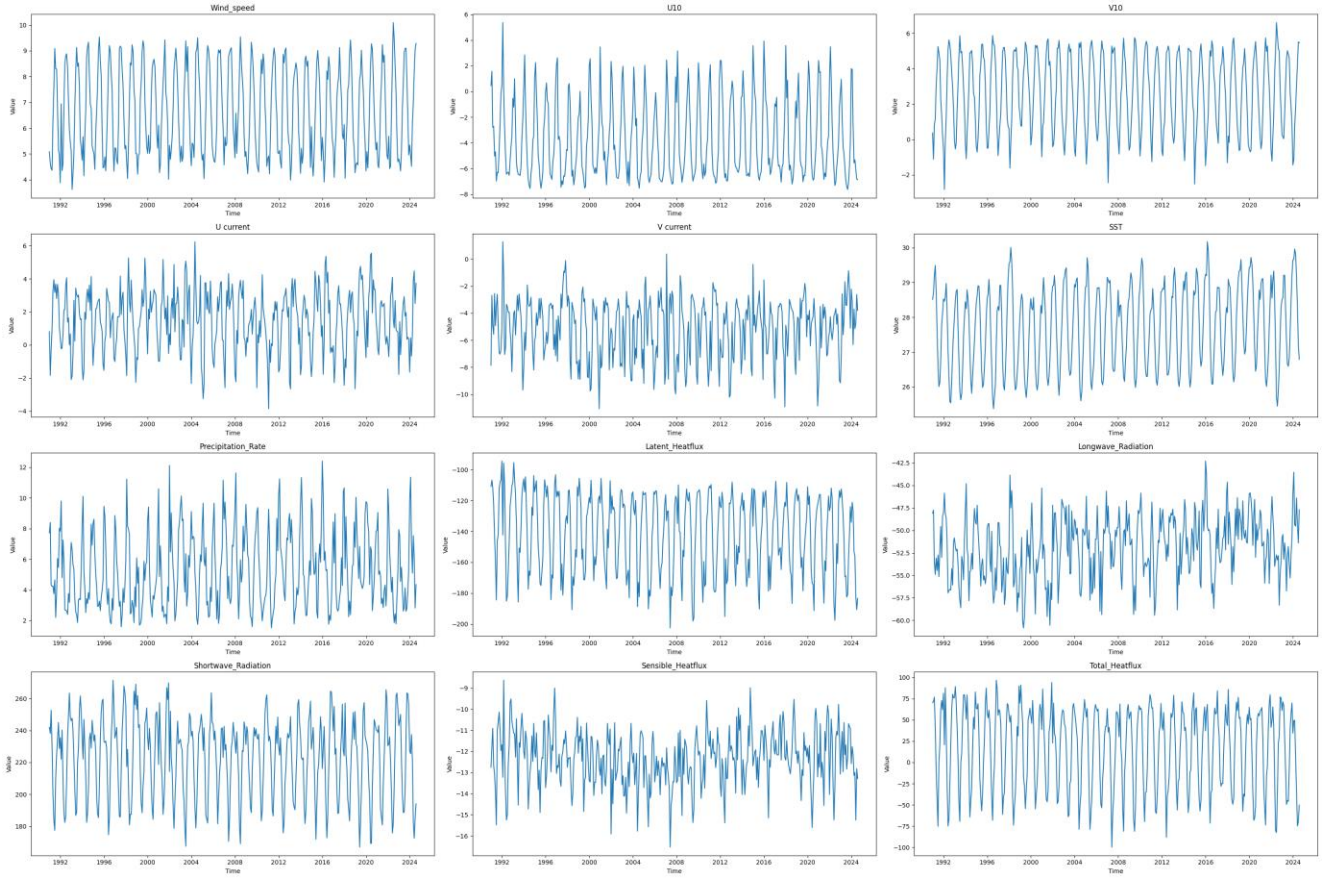
- 파생변수를 생성하여 에크만 수송과 해수면 열교환을 설명하고자 하였다
  - ➢ (1) U current = 50°E V wind – 80°E V wind => **zonal ocean current**
  - ➢ (2) V current = 15°S U wind – 5°S U wind => **meridional ocean current**
  - ➢ (3) Total heat flux = Solar + Thermal + Sensible + Latent heat flux

2 EDA

● 대기 및 해양 변수

- ERA5 monthly averaged data와 그 파생변수 총 12개를 수온약층 깊이 지수 'D20'과 비교
  - ➢ **Annual and seasonal variability**
  - ➢ **D20은 연간 변동성 외에도 큰 변동성이 존재**

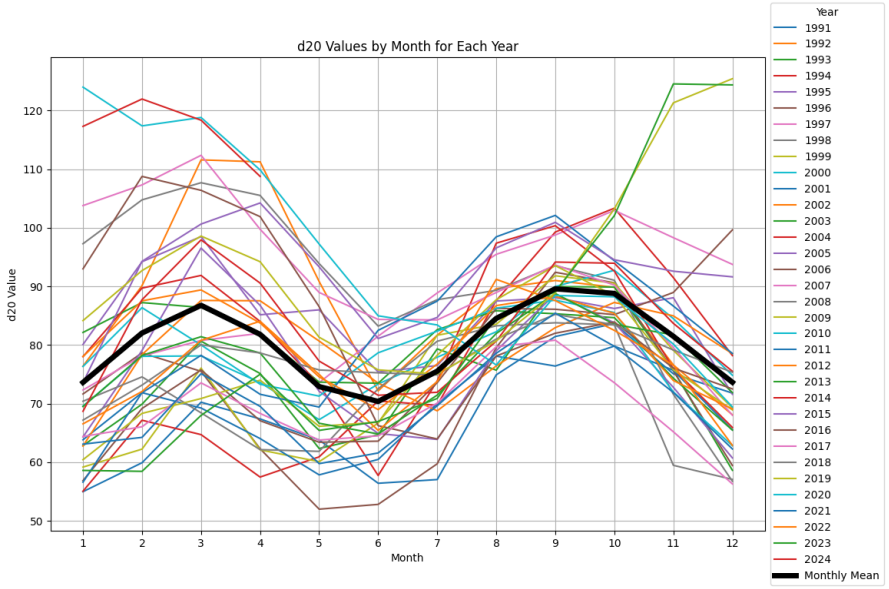


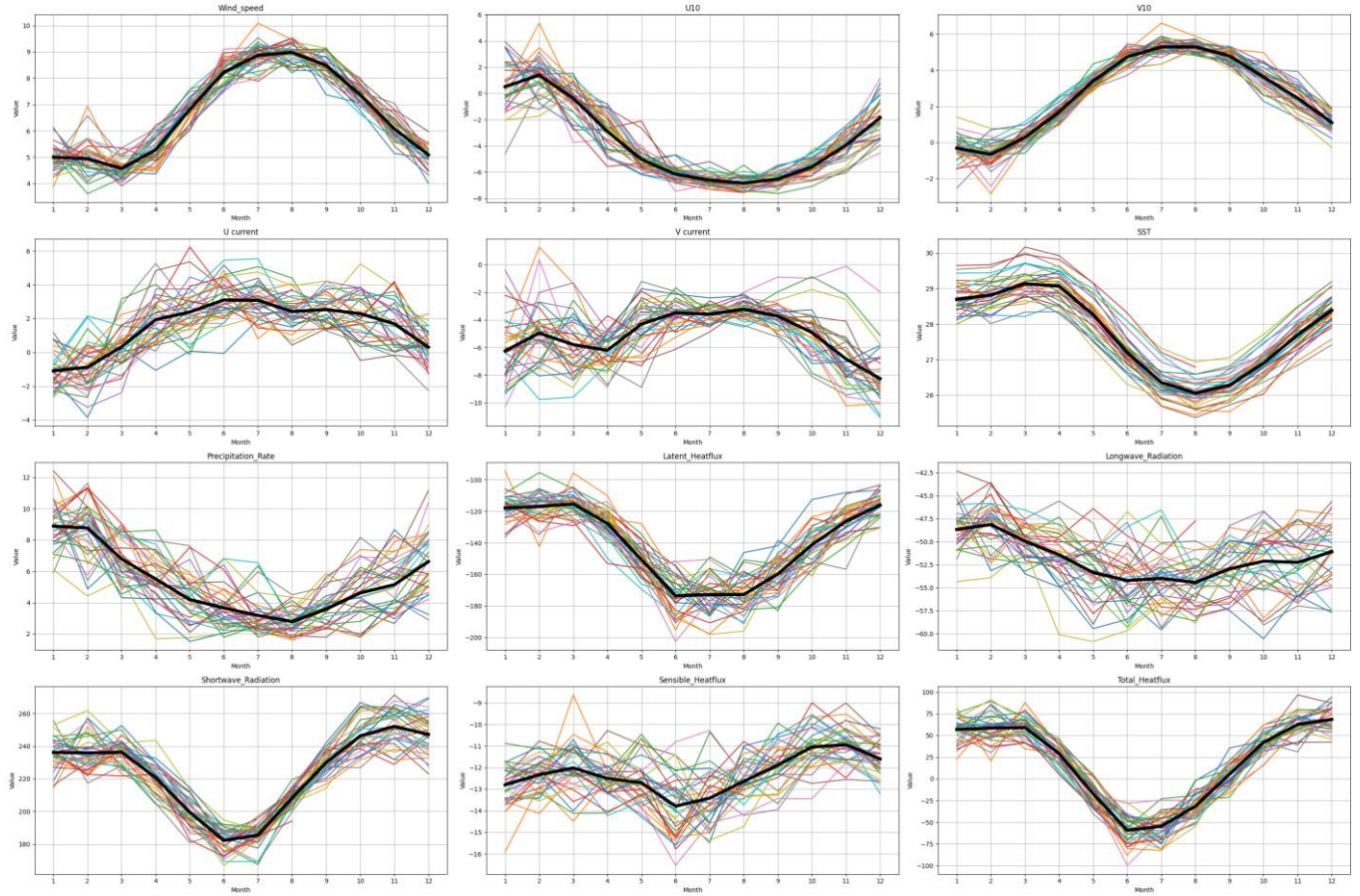


2 EDA

● 대기 및 해양 변수

▪ Annual and seasonal variability





- ➢ Dual-peaked pattern, Large variance
- ➢ 봄/가을에 최대, 여름/겨울에 최소
- ➢ Wind와 Heat flux가 주요하게 작용

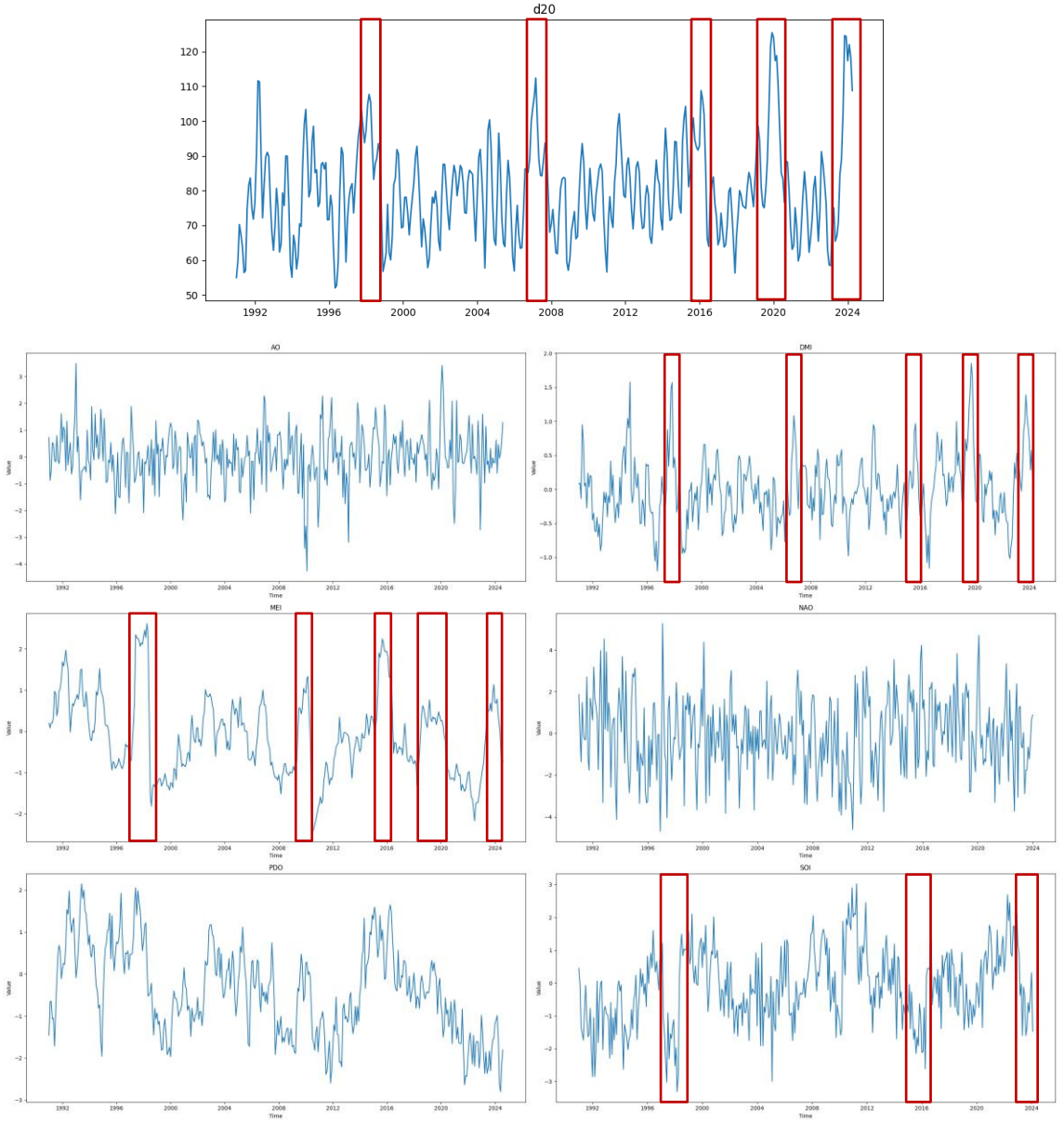
2 EDA

● 기후변동성 지수

- SCTR 지역과 6개의 기후변동성 간의 Teleconnection을 살펴보고자 하였다
  - ➢ 열대 기후변동성인 MEI, DMI, SOI과 북반구 기후변동성인 AO, NAO, PDO를 채택

- 선행연구에 따르면, SCTR 지역은 IOD, ENSO 변동과 연관 되어있다 (Vialard et al., 2009)

- 시계열 그래프를 살펴보면, D20의 주요 peak와 MEI, DMI, SOI의 peak가 유사한 시기에 발생했다



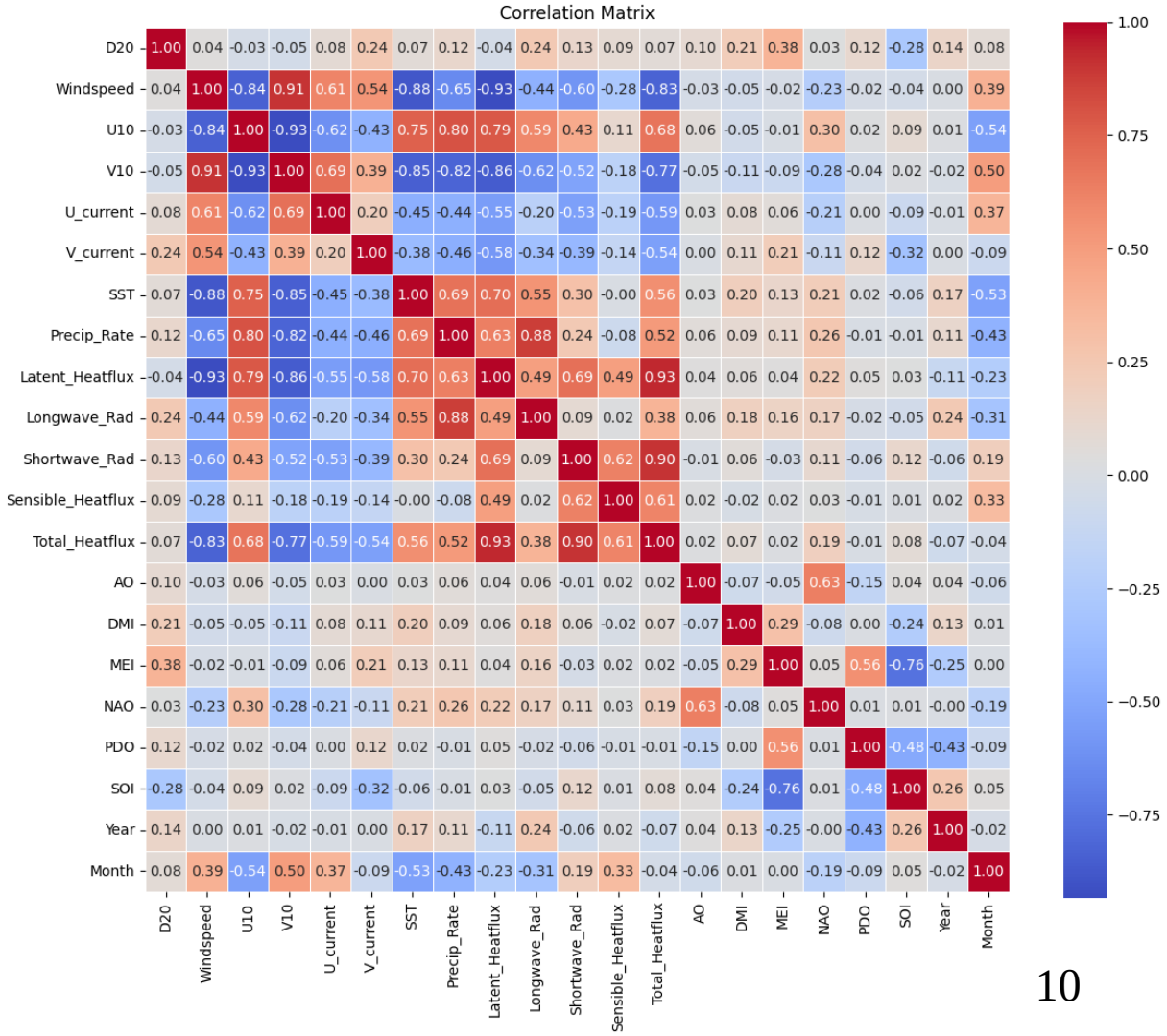
● 기후변동성 지수

- Multivariate ENSO Index (MEI)
  - ➢ **열대 태평양**에서 SST와 5개 기상변수를 EOF를 이용하여 만든 지수
- Dipole Mode Index (DMI)
  - ➢ **열대 동인도양과 서인도양**에서 SST 편차로, ENSO와 강한 상관관계가 있다고 알려져 있다
- Southern Oscillation Index (SOI)
  - ➢ **남동태평양과 서태평양**에서의 해수면 기압의 진동 패턴으로, 타히티와 다윈의 기압 차이로 정의한다
- Arctic Oscillation (AO)
  - ➢ 20°N 북쪽의 1000mb 높이의 1차 EOF로 정의하며, **북극 지역**의 해수면 기압의 강도를 나타낸다
- North Atlantic Oscillation (NAO)
  - ➢ 아조레스 고기압과 아이슬란드 저기압 간의 해수면 기압 차이를 기반으로, **북대서양**의 기후 변동을 나타낸다
- Pacific Decadal Oscillation (PDO)
  - ➢ 20°N-70°N 사이의 북태평양에서 월별 SSTA의 1차 EOF로 정의하며, **북태평양** SST 기후패턴을 나타낸다

3 Feature engineering

● 1차 변수 선택 : 상관관계

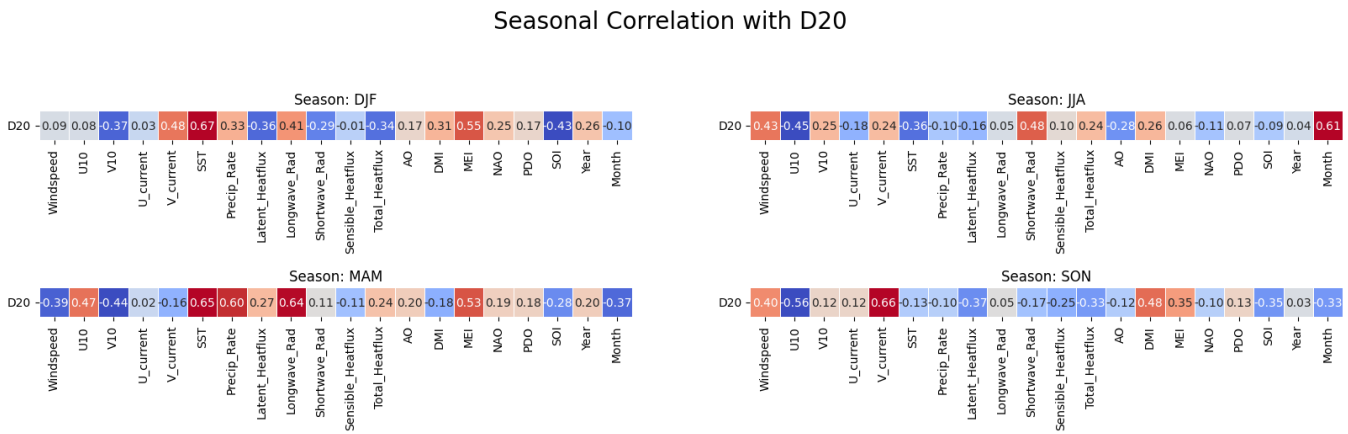
- D20과의 상관관계를 중심으로 변수 선택을 진행하였다
- D20과 다른 변수 간의 상관관계는 MEI를 제외하고는 매우 작아 상관성을 찾아보기 어렵다
- **Seasonality와 Lagged correlation**으로 인해 D20과의 상관성이 나타나지 않은 것으로 보인다

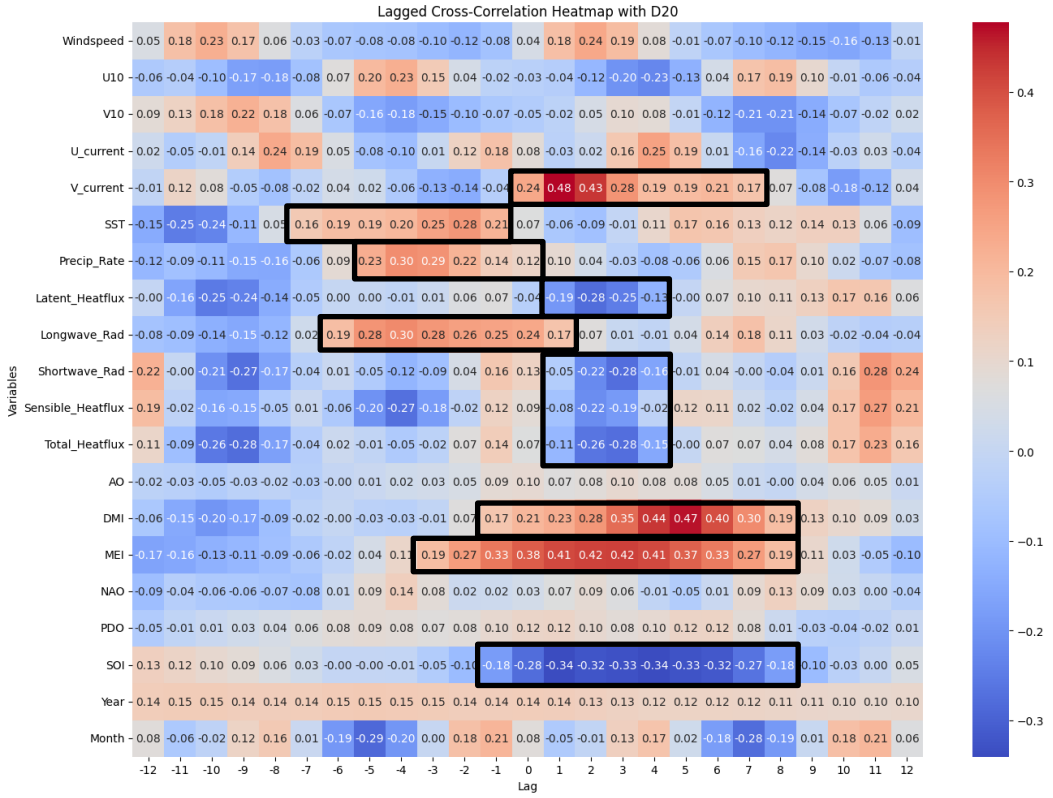


3 Feature engineering

● 1차 변수 선택 : 상관관계

- 1차 변수 선택: **Wind speed, V current, SST, Precipitation rate, Longwave Rad, Shortwave Rad, Total heat flux, DMI, MEI, SOI**
  - ➢ Season과 Lag를 고려한 상관관계를 통해 1차 변수 선택



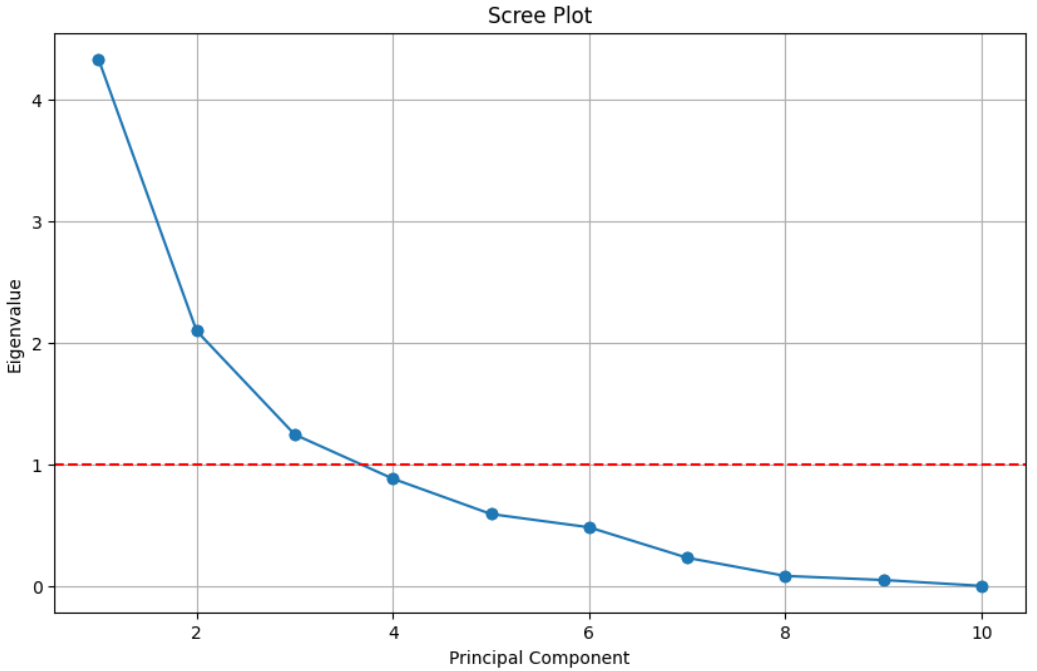


3 Feature engineering

● 2차 변수 선택 : 상관관계

- 1차 변수 선택으로 결정한 10개 변수에 **PCA를 진행**했다
  - ➢ PCA로 차원을 축소하면 모델의 복잡도를 낮춰, 예측 모델의 검증 성능을 강건화 할 수 있다.

- **Eigenvalue 값이** 1 이상(Kaiser rule)인 3개의 주성분을 선택했다
  - ➢ 2차 변수 선택: **PC1, PC2, PC3**

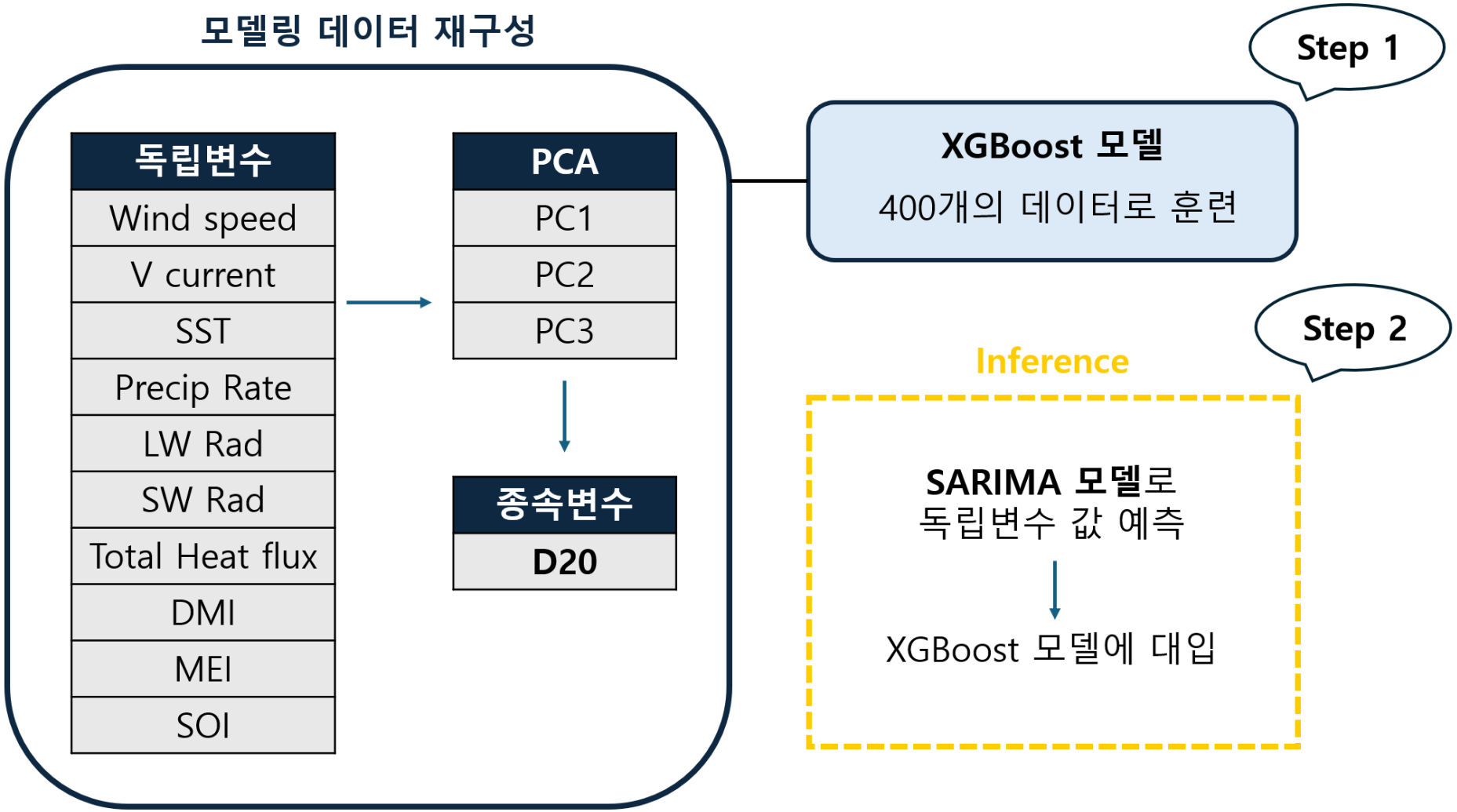


각 PC에 대한 X변수의 기여도

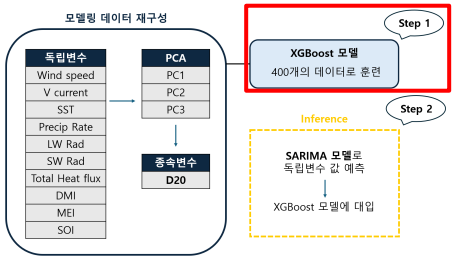
	Windspeed	V_current	SST	Precip_Rate	Longwave_Rad	Shortwave_Rad	Total_Heatflux	DMI	MEI	SOI
PC1	0.443074	0.322479	-0.395705	-0.394169	-0.330671	-0.311694	-0.421586	-0.042034	-0.012557	-0.039832
PC2	0.011095	0.251917	0.142142	0.140704	0.218007	-0.179347	-0.093554	0.326894	0.594147	-0.589998
PC3	-0.124080	0.092470	-0.113517	-0.386195	-0.471859	0.575238	0.371288	-0.010223	0.235025	-0.259374

4 Model

● Pipe Line



4 Model



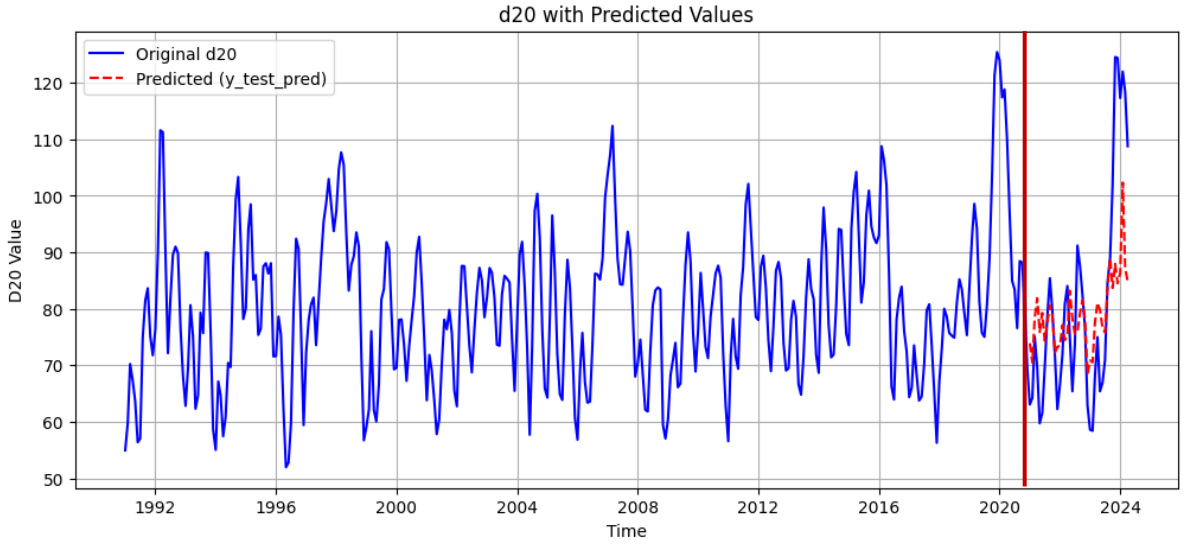
● STEP 1. XGBoost 모델

- 먼저 Train, Test를 9:1로 분리하고 Z-score scaling을 적용하였다
  - ➢ 시계열을 반영하기 위해 Train과 Test를 21년 1월을 기준으로 연속적으로 분리

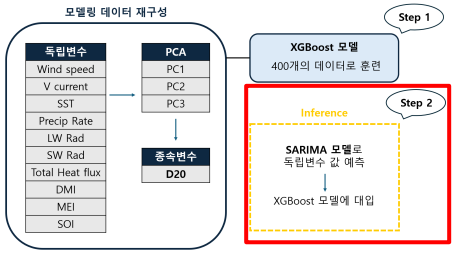
- **XGBoost Regressor**
  - ➢ 예측 성능과 학습 속도 모두 뛰어난 머신러닝 모델
  - ➢ Scoring은 MSE를 사용, Randomized Search로 튜닝

- 최종 모델 성능
  - ➢ Test RMSE: 14
  - ➢ Test MAE: 11

하이퍼파라미터	파라미터 후보	최종 선택
n_estimators	[100, 200, 300, 400, 500]	100
learning_rate	[0.05, 0.1, 0.2, 0.3, 0.5]	0.05
max_depth	[3, 4, 5, 6]	4
subsample	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	0.6
colsample_bytree	[0.6, 0.7, 0.8, 0.9, 1.0]	0.9



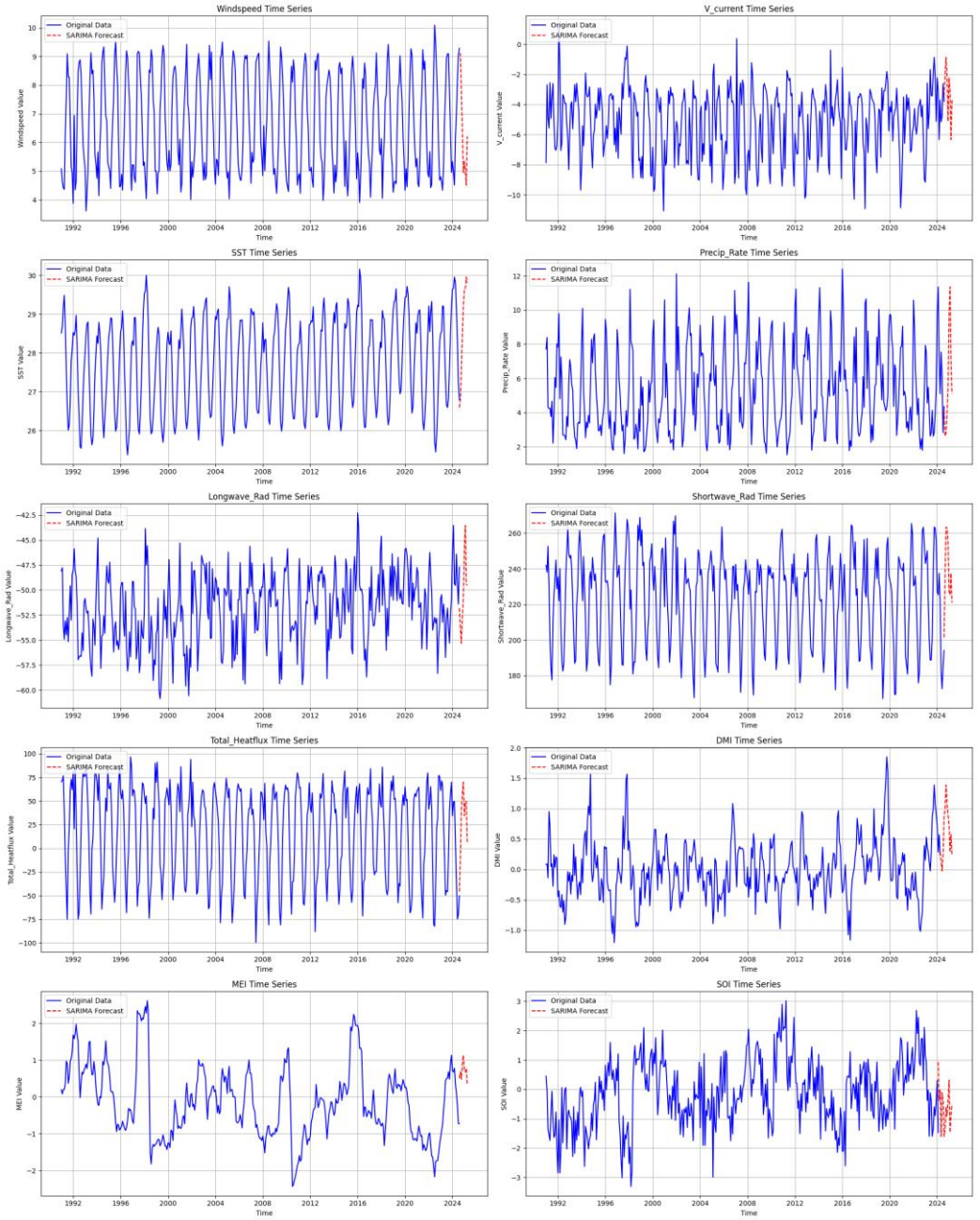
4 Model



● STEP 2. SARIMA 모델

- 2024년 5월부터 2025년 4월까지 SCTR 지역 D20을 예측하기 위해 해당 기간의 독립 변수가 필요하다
  - ➢ 각 변수별로 SARIMA 모델을 이용하여 예측

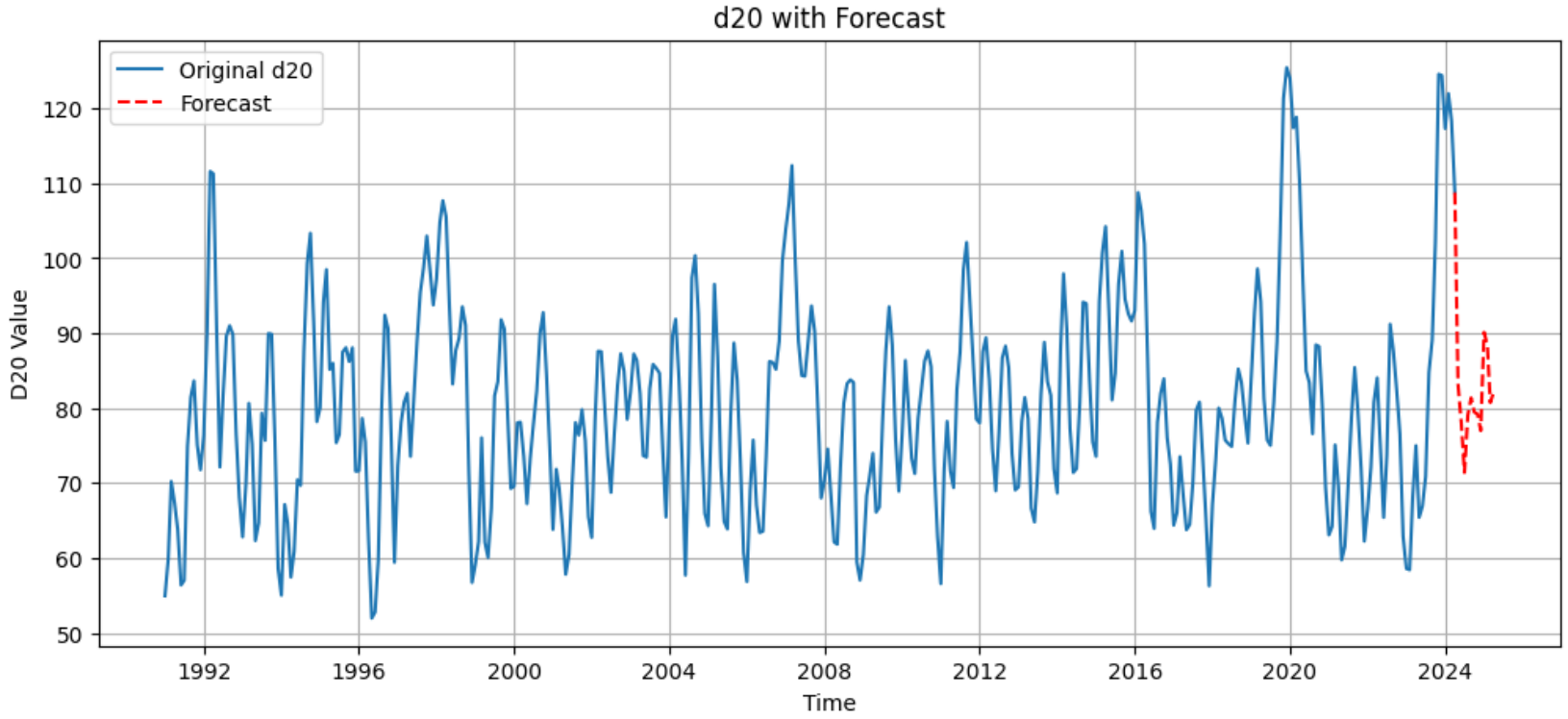
- **SARIMA model**
  - ➢ 계절적 변동성과 시계열을 고려한 통계 모델로, ARIMA 모델의 확장된 형태다
  - ➢ 주기를 12개월로 선택하여, 대기 및 해양 변수의 계절적 변동성을 반영한다
  - ➢ 오른쪽 그림을 보면 SARIMA로 예측한 각 독립변수를 확인 할 수 있다



5 Result

● Result

- 2024년 5월부터 2025년 4월에 대하여 **(1) SARIMA로 예측한 독립 변수를 (2) 사전 구축한 XGBoost 모델에 넣어 SCTR 지역 수온약층 깊이 지수를 Inference**했다



6 Disscusion

● 의의

- 인도양 및 전지구 해양 기후 예측 능력 강화
  - ➢ SCTR 지역의 수온약층 깊이는 IOD와 다양한 기후변동성과 연관 되어있다
  - ➢ 이 지역의 예측 능력을 향상함으로써 인도양 및 전지구 해양 기후 예측에 도움이 될 것으로 기대한다

- 머신러닝과 통계모델의 결합
  - ➢ SARIMA와 XGBoost를 활용한 예측모델은 시계열과 계절적 변동성을 반영한다
  - ➢ SARIMA와 XGBoost는 효율적인 계산을 통해 학습 속도가 빠르고 메모리 사용을 최소화한다

6 Disscusion

● 제언

- 데이터 크기의 한계
  - ➢ 학습 데이터가 월별 데이터 400개로 부족하여 LSTM, Transformer와 같은 딥러닝 모델을 사용할 수 없었다
  - ➢ **일별 데이터 또는 긴 기간의 학습 데이터로 딥러닝을 이용**하면 보다 좋은 성능을 기대할 수 있을 것이다

- 지구온난화로 인한 인위적인 강제력
  - ➢ 지구온난화로 기후 변동성이 커졌으며, 모델에서 **지구온난화의 패턴을 포함**하여야 더 정확한 미래 수온약층 변화를 예측할 수 있을 것이다

- 더 다양한 기후변동성
  - ➢ MJO나 QBO 등 **인도양에 영향을 미칠 수 있는 다양한 기후변동성**을 포함하면 예측 성능이 좋아질 수 있다

0 Reference

- 이 논문은 2022년 해양수산부 재원으로 한국해양과학기술진흥원의 지원을 받아 수행된 연구임(인도양 한-미 공동관측 및 연구, RS-2022-KS221662)

- Vijith, V. et al. (2020). Closing the sea surface mixed layer temperature budget from in situ observations alone: Operation Advection during BoBBLE. *Scientific Reports*, 10(1).1-12.
- Vialard, J. et al. (2009). CIRENE: Air–Sea Interactions in the Seychelles–Chagos Thermocline Ridge Region. *Bulletin of the American Meteorological Society*, 90(1), 45-62.