Check for updates
Journal of Korean Society for Atmospheric Environment
Vol. 41, No. 3, June 2025, pp. 430-447
https://doi.org/10.5572/KOSAE.2025.41.3.430
p-ISSN 1598-7132, e-ISSN 2383-5346

논 문

기계학습 기법을 활용한 WRF-CMAQ 모델의 $PM_{2.5}$ 구성성분 모의 성능 개선

Enhancing the Simulation Performance of $PM_{2.5}$ Compositions in the WRF-CMAQ Modeling System Using Machine Learning Techniques

김지민[1], 최민서[2], 전예지[1], 김태희[2], 곽경환[1),2),*], 이그림[3], 강병철[3], 정선아[3]

[1]강원대학교 환경학과, [2]강원대학교 미세먼지통합관리학과,
[3]국립환경과학원 대기환경연구부 대기환경연구과

접수일 2025년 3월 14일
수정일 2025년 4월 8일
채택일 2025년 4월 9일

Received 14 March 2025
Revised 8 April 2025
Accepted 9 April 2025

***Corresponding author**
Tel : +82-(0)33-250-8575
E-mail : khkwak@kangwon.ac.kr

Jimin Kim[1], Minseo Choi[2], Yeji Jeon[1], Taehee Kim[2], Kyung-Hwan Kwak[1),2),*], Greem Lee[3], Byeong-Cheol Kang[3], Sun-A Jung[3]

[1]*Department of Environmental Science, Kangwon National University, Chuncheon, Republic of Korea*
[2]*Department of Integrated Particulate Matter Management, Kangwon National University, Chuncheon, Republic of Korea*
[3]*Atmospheric Environmental Research Department, National Institute of Environmental Research (NIER), Incheon, Republic of Korea*

Abstract $PM_{2.5}$ compositions are important indicators for identifying emission sources and formation pathways of particulate matters in the atmosphere. In Korea, the Ministry of Environment has operated Air Quality Research Centers to monitor $PM_{2.5}$ components continuously. However, relying solely on measurement data has limitations on obtaining temporal and spatial information. A 3-D chemistry-transport model enables us to simulate $PM_{2.5}$ component concentrations at high spatio-temporal resolutions realistically when the simulated results are accurate. Therefore, this study aims to improve the simulation performance of one of the 3-D chemistry-transport models, Weather Research and Forecasting (WRF) - Community Multiscale Air Quality (CMAQ) model, for $PM_{2.5}$ and its components using machine learning techniques. The WRF-CMAQ simulation results, including $PM_{2.5}$ components, meteorology, geography, and emissions, were used as input data in the machine learning models. Measurement data of $PM_{2.5}$ and its components from Air Quality Research Centers at 10 locations were used as target variables to build the machine learning models. The study period was from January 1st to March 31st, 2022. The best machine learning model showed a correlation coefficient above 0.83 which is quite reasonable to use for $PM_{2.5}$ and its component simulations. We analyzed the WRF-CMAQ simulation results for $PM_{2.5}$ episodes occurred nationwide. The machine learning-corrected WRF-CMAQ model results captured nationwide high $PM_{2.5}$ levels better than the uncorrected WRF-CMAQ model results. It is expected that $PM_{2.5}$ characteristics in other regions where Air Quality Research Centers do not exist can be accurately provided using the machine learning-corrected WRF-CMAQ model.

Key words: WRF-CMAQ model, Performance enhancement, Machine learning, $PM_{2.5}$ components, Air Quality Research Center

1. 서 론

$PM_{2.5}$는 직경이 2.5 μm 이하인 입자로 탄소 성분(OC, EC), 질산염(NO_3^-), 황산염(SO_4^{2-}), 암모늄염(NH_4^+) 등 다양한 화학 성분의 혼합 물질이다. 이온 성분인 황산염, 질산염, 암모늄염은 대기에서 2차적으로 생성되는 물질로 배출원을 파악하는 데 중요한 역할을 한다(Zhang *et al.*, 2022; Kang *et al,* 2021; Nava *et*

al., 2020). 우리나라의 일평균 $PM_{2.5}$ 농도는 겨울철에 해당하는 12월부터 이듬해 3월까지 대기환경기준을 빈번하게 초과한다. 따라서, 겨울철 고농도 $PM_{2.5}$를 관리하기 위해 환경부에서는 매년 12월부터 3월까지 $PM_{2.5}$ 및 전구물질의 배출을 규제하는 미세먼지 계절관리제를 시행하고 있다(MOE, 2022). 이와 함께, 대기관리권역을 설정하여 권역마다 상이한 미세먼지 발생 원인에 맞추어 권역별로 대응하고 있다.

최근 미세먼지 농도가 점차 감소하는 추세를 보임에 따라 미세먼지 농도 저감과 더불어 대기 중 입자상 물질의 건강 위해성을 관리하는 정책으로의 전환이 요구되고 있다. 현행 미세먼지 관리 정책은 주로 총 미세먼지 질량 농도를 달성 목표로 설정하고 있으나, 총 미세먼지 질량 농도 기준의 규제만으로는 건강 영향을 효과적으로 관리하기 어렵다. 또한 미세먼지 질량 농도만으로는 미세먼지의 배출 특성이나 대기 중 생성, 소멸 과정 등을 파악할 수 없다는 점도 장애 요인으로 작용한다(Bae *et al.*, 2020; Lee *et al.*, 2015). 미세먼지의 구성성분 농도는 건강 위해성 등의 정보를 제공함으로써 미세먼지 관리 정책의 우선 순위를 도출하는 근거가 된다(Park *et al.*, 2010). 이처럼 미세먼지 구성성분 농도 자료의 중요성이 높아지면서 국립환경과학원 대기환경연구소는 미세먼지 구성성분을 포함한 대기오염물질 농도 자료를 실시간으로 생산하고 있다. 2008년 백령도 대기환경연구소를 시작으로 2025년 기준 총 12개소가 개소하였으며, 미세먼지 구성성분 등 주요 대기오염물질의 농도를 상시 측정하고 있다.

한편, 환경정책의 효과성을 평가하기 위한 목적으로 시공간적 제약이 적은 중규모 화학수송모델이 널리 활용되고 있다. 중규모 화학수송모델은 배출원과 수용원 간 관계를 규명하고 미세먼지 성분 농도별 저감 효과를 정량적으로 파악하는 데 효과적이다. Community Multiscale Air Quality (CMAQ) 모델은 미국 환경보호청(United States Environmental Protection Agency, US EPA)에서 개발한 중규모 화학수송모델로 미세먼지의 성분 농도를 모의하는 데 널리 사용되고 있다. Kang *et al.* (2021)은 CMAQ 모델을 활용해 $PM_{2.5}$ 성분 농도의 지역별 차이를 분석하고 지역별 배출량 관리의 중요성을 확인한 바 있다. CMAQ 모델은 연속적인 시공간적 모의가 가능하다는 장점이 있지만, 기상 및 배출 입력자료에 기인한 불확실성이 모의 성능을 낮춘다고 알려져 있다(Cha *et al.*, 2023; Bae *et al.*, 2023; Liu *et al.*, 2022; Wu *et al.*, 2021). 한 예로, Choi *et al.* (2019)은 CMAQ 모델로 동아시아 지역의 $PM_{2.5}$ 성분 농도를 모의한 결과에서 기상 조건의 불확실성으로 최대 55%의 오차가 발생했다고 보고하였다. 이러한 모의 농도와 관측 농도 간 차이는 CMAQ 모델을 이용하여 관심 지역의 미세먼지 발생 기작과 정책 효과성을 분석하는 데 장애물로 작용하고 있다.

CMAQ 모델의 모의 정확도를 높이기 위해서는 기상 및 배출 입력자료에 기인한 불확실성을 먼저 해소할 필요가 있지만, 이를 위해 수치 모델의 물리화학 과정을 개선하는 작업은 단순하지 않고 모든 지역과 기간에 걸쳐 일반화하기가 어렵다. 이러한 한계를 극복하기 위한 대안 중 하나로 기계학습 모델을 활용한 모의 성능 개선이 있다(Huang *et al.*, 2023; Kim *et al.*, 2023). 기계학습 모델은 데이터의 패턴과 구조를 학습하여 예측, 군집화 등을 수행하는 기술이다. 기상 변수, 대기오염물질 농도, 배출량, 지리적 특성 등 $PM_{2.5}$의 생성과 관련된 다양한 변수들이 $PM_{2.5}$ 농도에 미치는 상호작용을 학습하게 되며, 이를 통해 수치 모델은 기계학습 모델의 도움을 받아 더 정확한 모의 결과를 산출할 수 있다. 따라서, 최근에는 기계학습 모델과 수치모델을 결합하여 모의 성능을 높이는 연구가 활발히 진행되고 있다. 선행 연구에서는 기계학습 모델인 Random Forest (RF) 알고리즘을 통해 오존 농도와 영향 요인 사이의 비선형적 관계를 추정하거나 RF를 접목한 새로운 모델을 개발하여 CMAQ 모델의 예측 편향을 감소시킨 바 있다(Xiong *et al.*, 2024; Thongthammachart *et al.*, 2021). Do *et al.* (2023)은 DNN (Deep Neural Network) 모델을 활용하여 CMAQ 모델의 $PM_{2.5}$ 모의 정확도를 향상시킬 수 있음을 확인하였다. CMAQ 모델의 $PM_{2.5}$ 모의 농도, 기상자료, 대기오

염 측정 자료를 활용해 DNN 모델이 오차 패턴을 학습함으로써 기존 CMAQ 모델의 한계를 보완하여 미세먼지 발생 원인 분석의 정확도를 높였다. 이처럼 기계학습 모델을 활용하면 CMAQ 모델에서 불확실성을 유발하는 기상 조건과 배출 특성 등을 고려하여 격자별 농도를 보정하고 이를 바탕으로 해당 지역의 대기오염물질 생성 원인을 이해할 수 있다.

기계학습 기법은 수치 모델의 $PM_{2.5}$ 농도 공간 분포를 개선함으로써 미세먼지 발생 원인 분석 및 예측 성능 향상에 도움을 준다(Kim and Lee, 2023). Kim *et al.* (2023)은 대기질 패턴에 따른 대기오염물질 예측 성능 향상을 위해 지역 맞춤형 대기오염물질 예측 모델을 구축하였으며, 군집별 주요 오염원의 차이를 반영하여 지역별 미세먼지 발생 특성을 분석할 수 있는 기반을 마련한 바 있다. 그러나 앞서 언급한 선행 연구들은 주로 총 $PM_{2.5}$ 질량 농도 모의 성능 개선에만 초점을 맞췄다는 한계가 있다. 반면, $PM_{2.5}$ 성분 농도는 그 중요성에도 불구하고 관측 지점 수가 많지 않아서 기계학습 기법을 적용하는 데 제약이 있었다. 한편, 최근 국립환경과학원 대기환경연구소 관측 지점이 확대되면서 $PM_{2.5}$ 성분 농도의 관측자료가 빠르게 축적되고 있다. 이를 기계학습 모델의 학습자료로 활용한다면 CMAQ 모델에서 총 $PM_{2.5}$ 질량 농도뿐만 아니라 $PM_{2.5}$ 성분 농도의 모의 성능 향상도 기대할 수 있다.

본 연구의 목적은 기계학습 모델이 미세먼지 구성성분을 연속 측정하고 있는 국립환경과학원 대기환경연구소의 관측자료를 학습함으로써 CMAQ 모델의 $PM_{2.5}$ 구성성분 농도 모의 성능을 향상시키는 것이다. 특히, 국내 미세먼지 고농도 기간인 1~3월을 대상으로 해당 기간에 최적화된 기법을 제안하고자 하며, 대기환경연구소를 운영하지 않는 지역에도 신뢰성 있는 CMAQ 모델 자료를 제공하는 것이 목표다.

2. 연구 방법

2.1 연구 지역 및 대기환경연구소 자료

모의 대상 영역은 국외 수송을 고려하기 위한 동아시아 영역(27 km 격자)과 한반도 영역(9 km 격자), 권역 간 수송의 영향을 상세히 보기 위한 남한 영역(3 km 격자) 등으로 구성하였다. 연구 대상 기간은 미세먼지 고농도 기간을 포함하는 2022년 1월 1일부터 3월 31일이다. 환경부 국립환경과학원은 권역별 미세먼지 발생 원인 규명을 위해 권역별 대기환경연구소를 확대 운영하고 있으며, 본 연구에 사용한 대기환경연구소 자료는 2022년 기준으로 운영 중인 백령도, 수도권(서울), 호남권(광주), 중부권(대전), 제주도, 영남권(울산), 경기권(안산), 충청권(서산), 전북권(익산), 강원권(춘천) 등 총 10개소를 포함한다(그림 1).

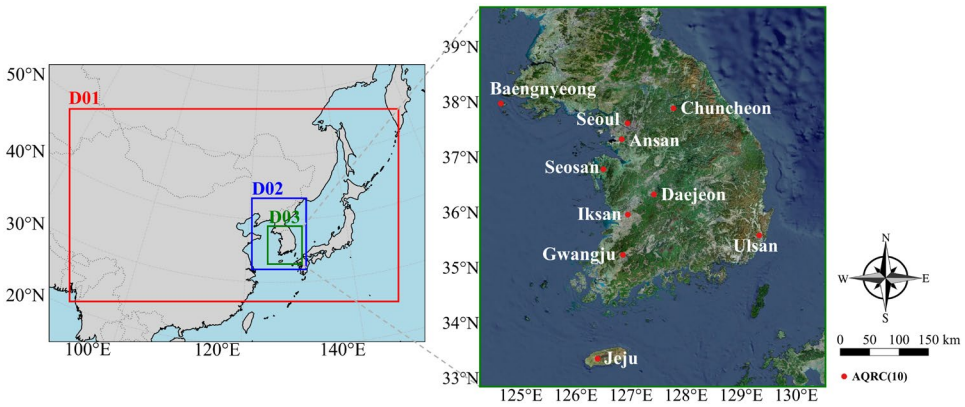

Fig. 1. Simulation domains of CMAQ model (left). Locations of Air Quality Research Center (AQRC) in South Korea (right).

권역별 대기환경연구소에서 생산되는 자료 중 1시간 간격으로 제공되는 $PM_{2.5}$ 질량 농도, 이온성분(NO_3^-, SO_4^{2-}, NH_4^+, Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-) 농도, 탄소성분(OC, EC) 농도를 사용하였다.

2021년 대기환경연구소 연간 운영 결과보고서에 따르면(NIER, 2022), $PM_{2.5}$ 질량 농도는 베타선 흡수법의 원리를 이용한 BAM-1020 (Met One Ins.)으로 측정한다. 대기 중 입경 2.5 μm 이하의 입자를 16.7 L/min 유속으로 필터에 채취한 후 시료 채취 전과 후의 베타선 감쇠량으로부터 $PM_{2.5}$의 질량 농도를 계산하는 방식이다. 이온성분 농도는 Ambient Ion Monitor (URG CO., 9000D, AIM)로 측정한다. 대기 중 에어로졸을 3 L/min의 유량으로 채취하며, 채취된 시료를 이온크로마토그래피(Ion Chromatography, IC) 기법으로 분석한다. 탄소성분 농도는 NIOSH 분석법 및 EPA STN 분석법을 기반으로 한 열광학적 분석법(thermal/optical transmittance method)과 비분산 적외선 분석법(non-dispersive infrared method)으로 측정한다.

2.2 WRF-CMAQ 모델

CMAQ 모델은 기상장과 배출량 입력자료를 필요로 한다. 기상 입력자료는 중규모 기상 모델인 Weather Research and Forecasting model (WRF) version 4.3을 사용하였다. WRF는 96시간 간격으로 모의하였으며, 모델의 초기 안정화 시간으로 24시간은 분석에서 제외하였다. CMAQ 모델은 14일의 초기 안정화 이후 24시간씩 연속 모의를 수행하였다. 모의 대상 기간은 CMAQ 모델의 안정화 시간을 포함하여 2021년 12월 17일 00 UTC부터 2022년 4월 1일 00 UTC까지다. 기상 모델의 초기 및 경계 조건은 National Center for Environmental Prediction (NCEP)에서 제공하는 Final Analysis (FNL) 1.0° × 1.0° 자료를 사용하였다. Land-use/land-cover (LULC) 입력자료로 27 km와 9 km 영역에는 United States Geological Survey (USGS)의 제공 자료를 사용하였으며, 3 km 영역에는 2020년 우리나라 환경부(Korea Ministry of Environment, KME)에서 제공한 최신 토지피복 자료를 사용하였다. WRF 모델이 생산한 기상자료는 the Meteorology-Chemistry Interface Processor (MCIP)를 통해 전처리하여 CMAQ 모델의 입력자료로 사용되었다. CMAQ 모델에서 사용한 배출량 자료는 환경부 국가미세먼지정보센터에서 제공하는 27 km - 9 km - 3 km 영역 배출 입력자료

Table 1. Specific information on WRF-CMAQ simulation.

Model domain		D01	D02	D03
Horizontal resolution		27 km	9 km	3 km
Grid numbers	WRF	181 × 143 × 39	109 × 148 × 39	208 × 226 × 39
	CMAQ	174 × 128 × 38	99 × 138 × 38	198 × 216 × 38
WRF v4.3 option				
Initial/boundary data		FNL 1.0 × 1.0°		
SST		OISST/FNL		
Land surface model		Noah LSM	(Chen and Duhia, 2001)	
PBL scheme		YSU scheme	(Hong et al., 2006)	
LULC		USGS/MODIS/KME		
CMAQ v5.3.2 option				
Chemical option		SAPRC07	(Carter, 2010)	
Aerosol option		AERO6	(Appel et al., 2021)	
Emission inventory		CAPSS 2021/SIJAQ v2		
Advection scheme		Yamo	(Yamartino, 1993)	
Horizontal diffusion		Multiscale	(Louis, 1979)	
Vertical diffusion		ACM2	(Pleim, 2007)	

다. 이 중에서 인위적 배출량은 Sparse Matrix Operator Kernel Emissions (SMOKE) 모델링의 결과다(Benjey *et al.*, 2001). 국내 인위적 배출량은 2021년 대기정책지원시스템(CAPSS 2021) 배출 목록에 기반하였으며(NAIR, 2023), 국외 인위적 배출량은 Satellite Integrated Joint Monitoring of Air Quality (SIJAQ) version 2 배출 목록을 활용하였다. 자연적 배출량은 Model of Emissions of Gases and Aerosols from Nature (MEGAN)의 결과를 이용하였다(Guenther *et al.*, 2006).

본 연구에서는 CMAQ version 5.3.2를 활용하여 대기오염물질 농도장을 모의하였다. CMAQ 모델의 모의 영역은 WRF 모델을 따라 동아시아, 한반도, 남한을 포함한 지역으로 중심 위경도는 126°E, 38°N이다. 수평 격자 크기는 각각 27 km, 9 km, 3 km이며, 연직층은 38층으로 대기 경계층 내 대기오염물질의 권역간 수송 과정을 상세히 모의하도록 설계하였다. 지도투영법은 Lambert conformal conic project 좌표계를 기반으로 하였다. 가스상 물질 화학 메커니즘은 Statewide Air Pollution Research Center (SAPRC) 07을 사용하였으며(Carter, 2010), 에어로졸 모듈은 AERO6 (the aerosol module version 6)를 사용하였다. 이외 모델에 설정된 옵션들은 표 1에 정리하였다.

2.3 데이터 전처리 및 후처리 과정

본 연구에서는 $PM_{2.5}$ 질량 농도와 개별 구성성분 농도마다 각각의 기계학습 모델을 구축하였으며, 이때 10개 지점의 대기환경연구소 측정 자료를 모두 학습에 사용하였다. 입력한 학습자료는 $PM_{2.5}$, 이온 성분(NO_3^-, SO_4^{2-}, NH_4^+, Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-), 탄소 성분(OC, EC) 농도다. 구축한 기계학습 모델에서 CMAQ 모델 자료를 입력자료로, 대기환경연구소 측정 자료를 목적 변수로 설정하였다. 기계학습 모델 구축 시 입력자료에 결측값이 포함되면 학습이 이루어지지 않는다. 따라서 한 가지 성분이라도 농도 데이터가 없는 시간대는 분석 대상에서 제외하였다. 기계학습 입력 변수로 사용하기 위해 WRF-CMAQ 모델의 3 km 영역 내 대기환경연구소가 위치한 지점의 격자 자료를 추출하였다. 기계학습 모델 구축 시 전체 입력자료를 랜덤하게 8:2로 분할하여 학습과 검증에 사용하였다(그림 2). 입력자료의 독립 변수 간 다중공선성이 있는 경우 이를 배제하기 위해 독립 변수 간 0.8 이상의 상관계수를 보이는 두 변수 중 변수 기여도가 높은 변수 하나만 사용하였다(Guyon and Elisseeff, 2003). 또한 풍향과 풍속 자료는 u, v 성분 바람으로 변환하였다. 최종적으로 기계학습에 사용된 변수를 표 2에 정리하였다.

CMAQ 모델의 $PM_{2.5}$ 성분 농도를 기계학습 모델로 후보정하는 방법에 관한 모식도를 그림 3에 나타내었다. 기계학습 모델은 관측자료가 존재하는 격자 지점에서 입력 변수만을 학습하여 구축되었다. 해당 모델은 학습 시 활용된 변수들을 기반으로 실행되므로 동일한 변수를 입력하면 보정된 값을 산출할 수 있다. 변

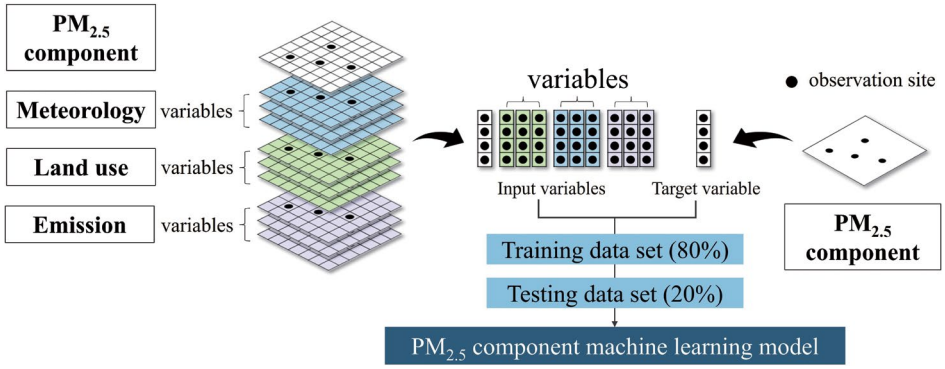

Fig. 2. Data pre-processing workflow in the machine learning model.

Table 2. Training data information used for machine learning model.

Category	Variables	Unit	Information
$PM_{2.5}$ component (Observation data)	$PM_{2.5}$	$\mu g/m^3$	$PM_{2.5}$ concentration
	NO_3^-	$\mu g/m^3$	Nitrate concentration
	SO_4^{2-}	$\mu g/m^3$	Sulfate concentration
	NH_4^+	$\mu g/m^3$	Ammonium concentration
	Na^+	$\mu g/m^3$	Sodium concentration
	K^+	$\mu g/m^3$	Potassium concentration
	Cl^-	$\mu g/m^3$	Chloride concentration
	Ca^{2+}	$\mu g/m^3$	Calcium concentration
	OC	$\mu g/m^3$	Organic carbon concentration
	EC	$\mu g/m^3$	Elemental carbon concentration
	Mg^{2+}	$\mu g/m^3$	Magnesium concentration
$PM_{2.5}$ component (CMAQ model data)	PM25_TOT	$\mu g/m^3$	$PM_{2.5}$ concentration
	PM25_NO3	$\mu g/m^3$	Nitrate concentration
	PM25_SO4	$\mu g/m^3$	Sulfate concentration
	PM25_NH4	$\mu g/m^3$	Ammonium concentration
	PM25_NA	$\mu g/m^3$	Sodium concentration
	PM25_K	$\mu g/m^3$	Potassium concentration
	PM25_Cl	$\mu g/m^3$	Chloride concentration
	PM25_CA	$\mu g/m^3$	Calcium concentration
	PM25_OC	$\mu g/m^3$	Organic carbon concentration
	PM25_EC	$\mu g/m^3$	Elemental carbon concentration
	PM25_Mg	$\mu g/m^3$	Magnesium concentration
Meteorological data	U10	m/s	U component of wind at 10 m
	V10	m/s	V component of wind at 10 m
	T2	K	Air temperature at 2 m
	RH2	%	Relative humidity at 2 m
	SLP	hPa	Sea level pressure
Land use data	LU_INDEX	1~33	Land use category
	HGT	m	Terrain elevation
	Latitude	°	Latitude
	Longitude	°	Longitude
Emission data	SO_2	moles/s	SO_2 emission
	NO_2	moles/s	NO_2 emission
	NH_3	moles/s	NH_3 emission
	Isoprene	moles/s	Isoprene emission
Time series data	Day	1~366	Day of year
	Hour	1~24	Hour of day

수 선정 시 CMAQ 모델의 성분 농도 외에 위경도, 토지 및 기상자료 등을 입력변수로 사용하였다. 이 중 위경도로 대표되는 위치 정보는 지상 관측 지점의 $PM_{2.5}$ 성분 농도를 학습한 후 전체 영역의 공간 분포를 개선하기 위한 입력자료로 사용된다(Di *et al.*, 2019). 따라서 기계학습 모델에 CMAQ 모의 결과를 입력하면 관측자료가 없는 지점에서도 기상 및 지형 조건을 반영하여 보정된 농도를 산출할 수 있다. 예를 들어, CMAQ 모델의 3 km 해상도 자료에서 개별 격자 내 특정 성분 농도를 기계학습 입력자료 형태로 변환 후 기계학습 모델에 입력하면, 해당 격자에서 보정된 성분농도를 얻게 된다. 이 과정을 CMAQ 모델의 모든 격자에 반복

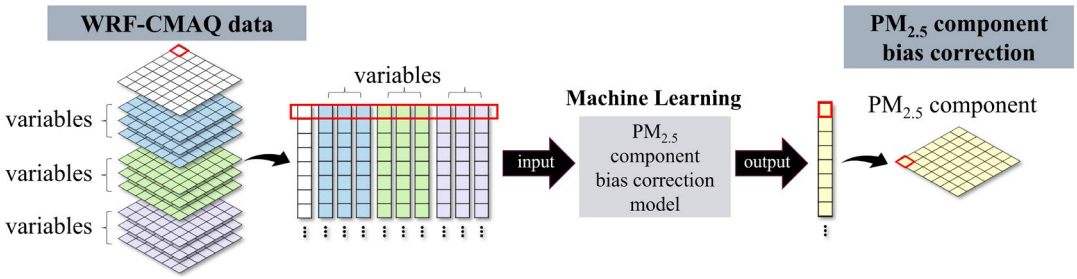

Fig. 3. Data post-processing workflow in the machine learning model.

적으로 수행하면 기계학습을 통해 보정된 CMAQ 모델의 $PM_{2.5}$ 구성성분 농도장을 생성할 수 있다.

2.4 모델 성능 평가 지표

본 연구에서는 검증지표로 상관계수(Correlation, r), 평균 제곱근 오차(Root Mean Square Error, RMSE), 평균 절대 오차(Mean Absolute Error, MAE)를 활용하였다. 위 지표들은 모델 예측 성능 평가에 주로 사용되는 지표이다. r 값은 1에 가까울수록 강한 양의 상관관계를 가지며, −1에 가까울수록 강한 음의 상관관계, 0에 가까울수록 상관관계가 약한 것을 의미한다. RMSE는 예측값과 관측값의 차이를 제곱한 후 평균을 구하고 그 평균값에 제곱근을 취하여 계산되며, 값이 작을수록 예측 모델의 성능이 정확함을 의미한다. MAE는 예측값과 관측값 간의 절대적인 차이를 평균하여 계산되며, 값이 작을수록 모델의 정확도가 높다. 검증지표는 식 (1)~(3)과 같이 계산되며 수식에서 M_i는 i번째 모델값, $\overline{M}$은 모델값의 평균, O_i는 i번째 관측값, $\overline{O}$는 관측값의 평균, N은 총 데이터 수를 의미한다.

$$r = \frac{\sum_{i=1}^{N}(M_i-\overline{M})(O_i-\overline{O})}{\sqrt{\sum_{i=1}^{N}(|M_i-\overline{M}|)^2}\sqrt{\sum_{i=1}^{N}(|O_i-\overline{O}|)^2}} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(M_i-O_i)^2} \quad (2)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(|M_i-O_i|) \quad (3)$$

3. 결과 및 고찰

3.1 최적의 알고리즘 선택

기계학습 알고리즘 중 트리 기반 앙상블 알고리즘(Tree-based ensemble algorithm)은 다양한 의사 결정나무(Decision Tree)를 결합하는 방법이다. 트리 기반 앙상블 알고리즘은 $PM_{2.5}$ 예측 연구에서 높은 예측 성능을 나타내는 것으로 보고되고 있다(Kim *et al.*, 2022; Ghahremanloo *et al.*, 2021; Park *et al.*, 2021). 기계학습은 입력자료 특성에 따라 최적의 알고리즘이 다르기 때문에 기계학습 모델 구축 시 알고리즘을 비교하여 최적의 알고리즘을 적용해야 한다(Kim *et al.*, 2023; Jeong and Kwak, 2022). 본 연구는 트리 기반 앙상블 알고리즘 중 Random Forest (RF), Extra Trees (ET), Extreme Gradient Boosting (XGboost), Light Gradient Boosting Machine (LightGBM)의 성능을 비교하였다. 최적의 알고리즘을 선정하기 위해 각 알고리즘에 대해 k겹 교차 검증(k-fold cross validation, k-fold)을 수행하였다. k-fold는 전체 데이터를 k개의 부분 집합으로 나눈 후 학습 집단과 검증 집단을 교차하여 k번 반복 검증하는 방법이며, 본 연구에서는 k를 5로 설정하였다. 교차검증 과정에서 r, RMSE, MAE와 같은 검증 지표를 사용하여 모델 예측 정확도를 평가한 후 가장 우수한 알고리즘을 최적의 알고리즘으로 선정하였다. 성분별 기계학습 모델의 검증지표를 알고리즘별로 평균한 결과, ET가 상관계수 0.92, MAE 0.78 μg/m^3, RMSE 1.27 μg/m^{3}로 가장 우수한 성능을 보였다. 중앙

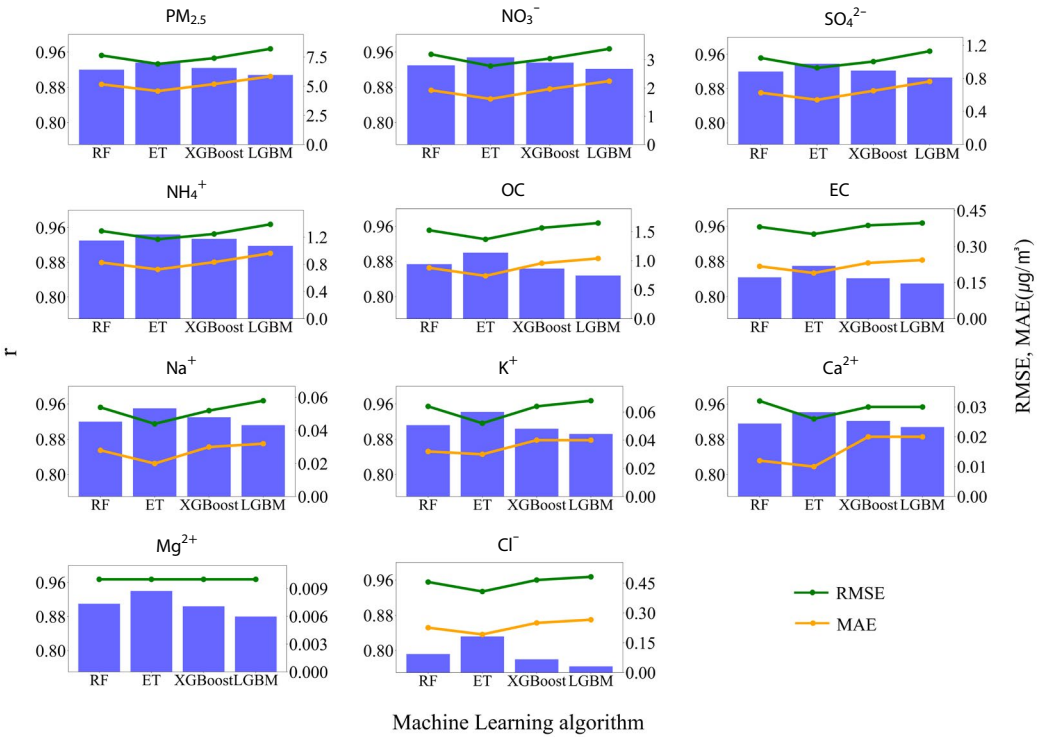

Fig. 4. 5-fold cross-validation results of machine learning models (RF, ET, XGBoost, and LGBM) for $PM_{2.5}$ components.

값 또한 ET가 상관계수 0.94, MAE 0.19 μg/m^3, RMSE 0.41 μg/m^{3}로 나타나 가장 우수한 성능을 나타냈다(그림 4). ET 알고리즘은 기존 RF 방법보다 의사결정 나무 분할 시 무작위성이 높아 데이터의 패턴을 학습하는 데 유리하기 때문에 우수한 예측 성능을 낼 수 있다(Geurts *et al.*, 2006). 따라서 본 연구에서는 ET를 최적 알고리즘으로 선정하여 기계학습 모델 구축에 사용하였다.

3.2 기계학습 모델 구축 및 검증

선정한 ET 모델을 이용하여 CMAQ 모델 자료와 대기환경연구소 측정 자료를 학습하여 모델을 구축하였다. 기계학습 모델 후처리 결과가 기존 CMAQ 모델 결과보다 얼마나 개선되었는지 평가하기 위해 대기환경연구소 측정 자료를 무작위로 나누어 전체 기간 중 80%는 학습, 20%는 검증에 사용하였다. 검증 자료를 통해 기존 CMAQ 모델의 $PM_{2.5}$ 및 구성성분 모의 농도와 측정 농도를 비교한 산점도를 그림 5에 나타냈다. CMAQ 모델 검증 결과에서 NO_3^-은 상관계수가 0.64, SO_4^{2-}는 0.62, NH_4^+는 0.70으로 나타났다. 이외 다른 이온 성분 Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-의 상관계수는 순서대로 각각 0.33, 0.35, 0.21, 0.39, 0.25였다. 탄소 성분인 OC, EC의 상관계수는 각각 0.43, 0.55로 나타났다.

그림 6은 그림 5와 마찬가지로 기계학습 검증 자료를 활용하여 기계학습으로 후처리된 CMAQ 모델(CMAQ-ML)의 $PM_{2.5}$ 및 구성성분 농도와 측정 농도를 비교한 산점도다. 기계학습 모델 검증 결과에서 NO_3^-은 상관계수가 0.95, SO_4^{2-}는 0.94, NH_4^+는 0.95로 성능이 우수하였다. 다른 이온 성분인 Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-의 상관계수는 순서대로 각각 0.96, 0.95, 0.95, 0.94, 0.83이었다. 탄소 성분인 OC, EC의 상

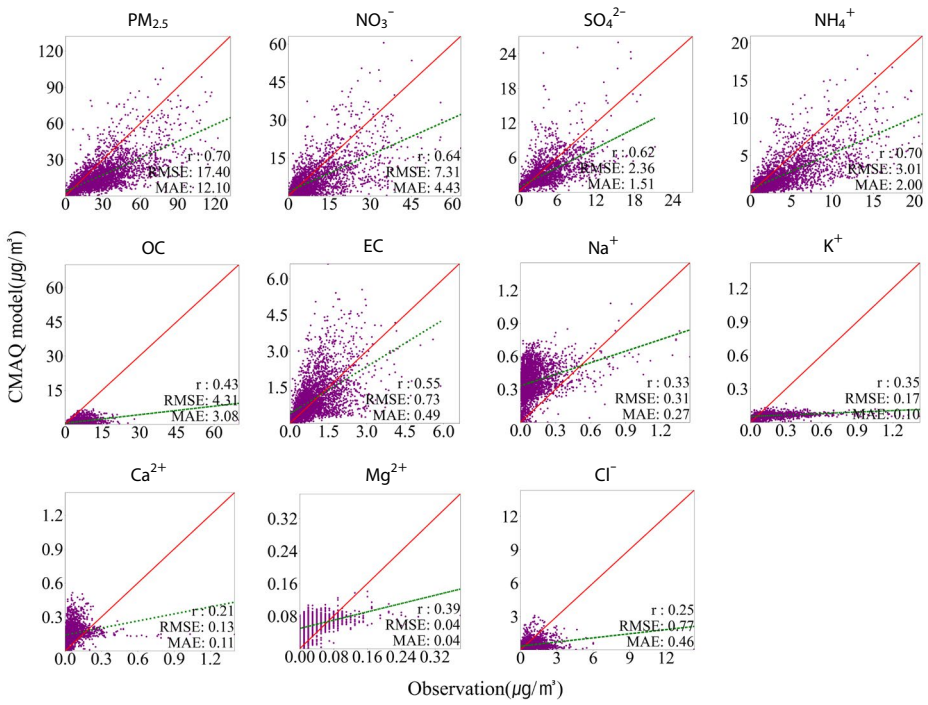

Fig. 5. Scatter plots comparing the simulated $PM_{2.5}$ and its component concentrations using the CMAQ model with the observed concentrations.

관계수는 각각 0.88, 0.89로 나타났다. 기계학습 모델을 적용하기 전의 기존 CMAQ 모델은 전반적으로 각 구성성분 농도를 과소 모의하는 경향을 보였으며, 특히 이온 성분 물질에 대한 모의 성능이 낮았다. 이는 기존 모델의 성분 농도에 대한 입력자료에 대한 불확실성과 함께 모델 내 황산염, 질산염, 암모니아 간 반응에 대한 메커니즘 이해가 부족하기 때문으로 판단된다(Smyth *et al.*, 2006). 반면, 기계학습 모델을 적용함으로써 전반적으로 구성성분 농도의 모의 성능이 개선됨을 확인하였다. 특히, 이온 성분에 대한 모의 성능이 향상되었으며, 탄소 성분의 상관계수도 개선되어 보다 현실적인 $PM_{2.5}$ 및 성분 농도를 모의하였다.

3.3 고농도 $PM_{2.5}$ 기간 적용 사례

고농도 $PM_{2.5}$ 대상 기간 선정은 미세먼지 비상저감조치 기준을 참고하였다. 미세먼지 비상저감조치란 당일 0시부터 16시까지의 평균 $PM_{2.5}$ 농도가 50 μg/m^{3}를 초과하고 다음 날의 일평균 $PM_{2.5}$ 농도가 50 μg/m^{3}를 초과할 것으로 예측되거나 또는 다음 날의 일평균 $PM_{2.5}$ 농도가 75 μg/m^{3}를 초과할 것으로 예측될 때 발령할 수 있다(MOE, 2022). 본 연구에서 위의 기준을 참고하여 연구 대상 기간 중 수도권을 기준으로 고농도 $PM_{2.5}$ 사례 기간을 선정하였다. 선정된 고농도 $PM_{2.5}$ 사례 기간은 2022년 2월 25일 10시부터 26일 11시까지다. 사례 기간의 평균 $PM_{2.5}$ 농도는 수도권 대기환경연구소에서 74.0 μg/m^{3}로 다른 권역보다 높았다. 경기권, 강원권, 호남권, 전북권, 중부권 대기환경연구소에서도 사례 기간 평균 $PM_{2.5}$ 농도가 60 μg/m^{3}를 초과하였다. 반면, 영남권, 제주도 대기환경연구소에서는 사례 기간 평균 $PM_{2.5}$ 농도가 50 μg/m^3 이하로 상대적으로 낮은 농도 수준을 보였다. 보도자료에 따르면 해당 기간 백령도에서 고농도 $PM_{2.5}$ 현상이 관측된 후 시차를 두고 내륙 지역에서 농도가 증가하는 경향을 보였다(Kwak, 2022). 또한 시베리아 고기압이 발

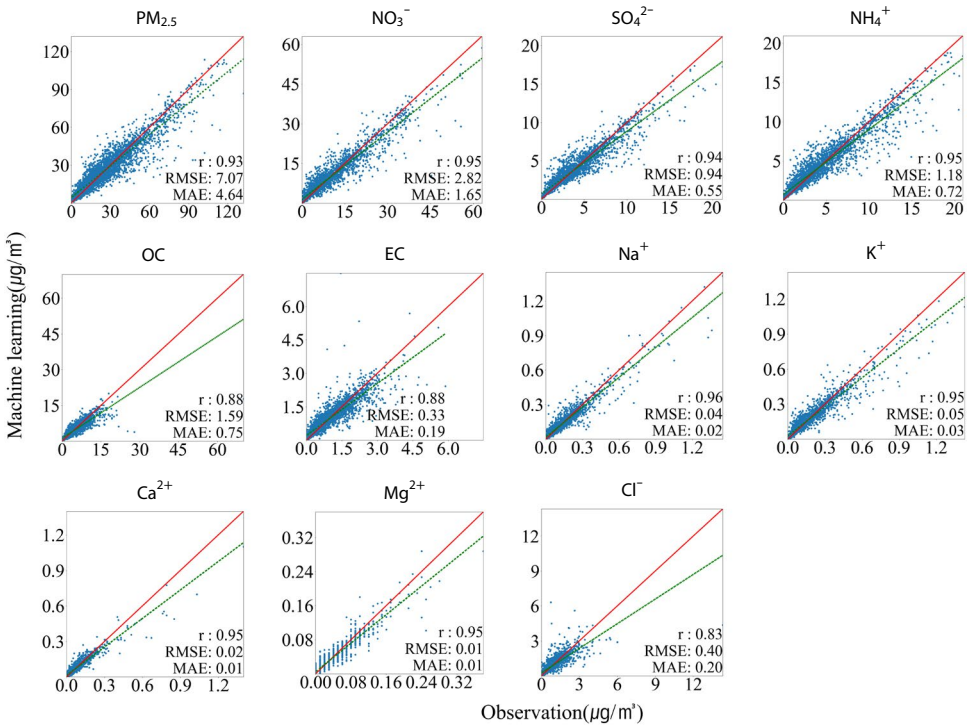

Fig. 6. Scatter plots comparing the simulated $PM_{2.5}$ and its component concentrations using the CMAQ-ML model with the observed concentrations.

달하며 북서풍이 강해진 것으로 보고되었다. 종합적으로 외부 유입에 의해 내륙 지역에서 고농도 $PM_{2.5}$ 현상이 발생한 것으로 판단된다(KMA, 2022).

그림 7은 고농도 $PM_{2.5}$ 사례 기간 동안 CMAQ 모델이 모의한 $PM_{2.5}$, NO_3^-, SO_4^{2-}, NH_4^+, OC, EC 농도의 분포와 대기환경연구소에서 관측된 성분 농도를 비교한 그림이다. CMAQ 모델이 모의한 $PM_{2.5}$의 공간분포를 분석한 결과 영남권에서 상대적으로 높은 농도가 나타났으며, 중부권과 남부권에서는 비교적 낮은 농도를 보였다. 제주도를 제외한 대부분의 권역에서 CMAQ 모델은 $PM_{2.5}$ 농도를 관측에 비해 과소 모의하는 경향을 보였으며, 이는 CMAQ 모델이 각 성분 농도를 전반적으로 과소 모의하였기 때문이다. 성분별 관측 농도 분포를 살펴보면 NO_3^-의 농도가 다른 성분보다 높게 나타나며, 특히 영남권역에서 높은 수준을 보였다. 그러나 CMAQ 모델은 NO_3^- 농도를 관측보다 과소 모의하는 것으로 나타났다. SO_4^{2-}는 관측 농도와 비교하여 수도권과 강원권 등 중부 지역에서 CMAQ 모델이 과대 모의한 반면, 남부 지역에서는 CMAQ 모델이 과소 모의하였다. NH_4^+, OC, EC의 모의 농도는 전국적으로 비슷한 수준을 보였다. 그러나 NH_4^+는 관측된 농도에 비해 CMAQ 모델이 과소 모의하였으며, OC는 지역별 관측 농도 차이가 뚜렷하였으나 모델은 그 경향을 모의하지 못하였다. 반면, EC는 관측 농도보다 과대 모의하는 경향을 보였다. 전반적으로 $PM_{2.5}$의 구성성분에서 큰 비중을 차지하는 NO_3^-와 NH_4^+를 과소 모의하는 경향이 강했다.

그림 8은 고농도 $PM_{2.5}$ 사례 기간 동안 CMAQ 모델이 모의한 $PM_{2.5}$, NO_3^-, SO_4^{2-}, NH_4^+, OC, EC 농도의 분포를 기계학습으로 후처리한 농도 분포 결과다. 기존 CMAQ 모델 결과는 $PM_{2.5}$ 농도를 전반적으로 과소 모의하였으나, 기계학습으로 후처리한 결과에서는 서

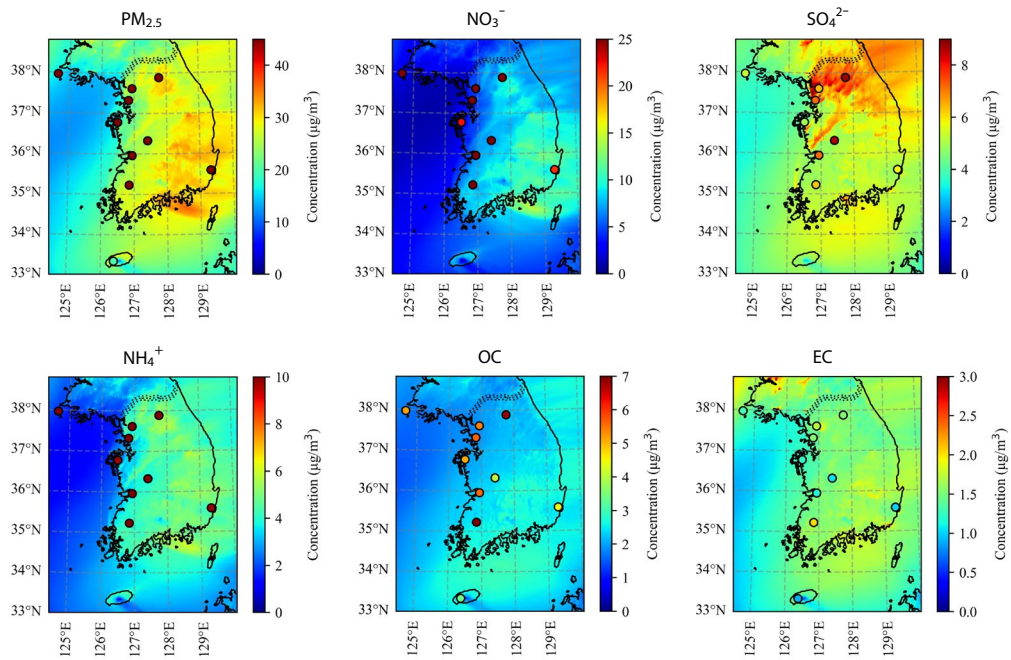

Fig. 7. The distribution of $PM_{2.5}$ and its components in the CMAQ model at the surface level averaged from 10 KST 25 to 11 KST 26 February, 2022. Colored circles indicate the observed concentrations.

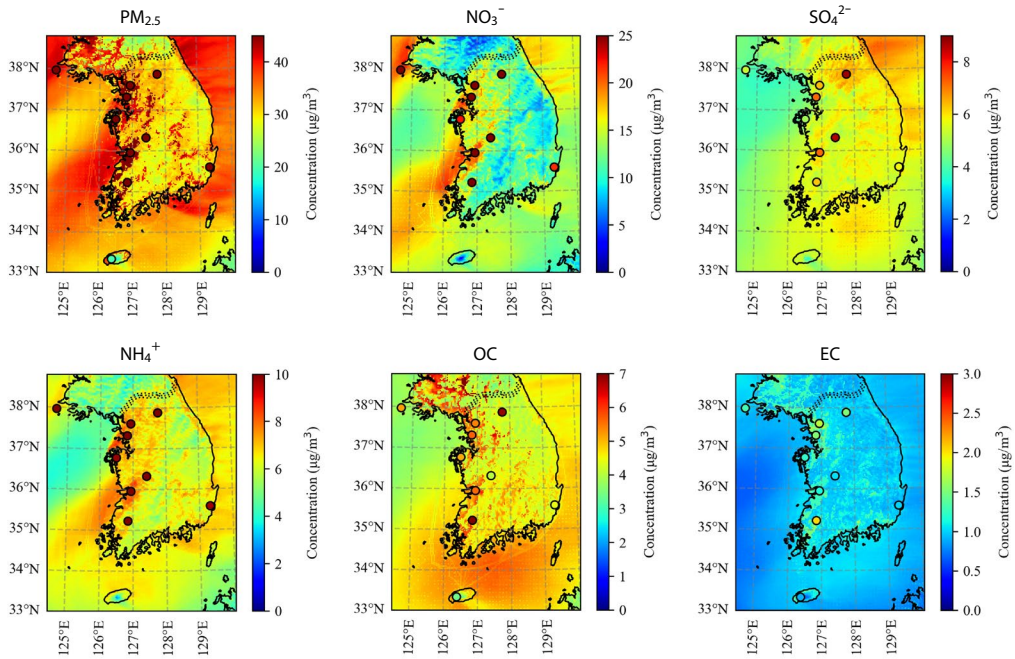

Fig. 8. The distribution of $PM_{2.5}$ and its components in the CMAQ-ML model at the surface level averaged from 10 KST 25 to 11 KST 26 February, 2022. Colored circles indicate the observed concentrations.

해안 지역과 수도권과 중부권에서 관측 농도와 유사한 수준으로 개선되었다. 이는 주로 NO_3^-, NH_4^+, OC와 같이 CMAQ이 과소 모의하던 성분의 농도가 수도권과 중부권에서 증가하였기 때문이다. 기존 CMAQ 모델은 전국적으로 비슷한 수준의 NH_4^+와 OC 농도를 보였으나, 기계학습으로 후처리하였을 때 서해안과 수도권에서 상대적으로 농도가 높아지면서 관측 농도와 유사해졌다. NO_3^-도 서해안과 수도권에서 농도 편차가 개선되었지만, 여전히 과소 모의하는 경향을 보였다. SO_4^{2-}는 기계학습으로 후처리하기 전 CMAQ 모델이 관측 농도와 비교하여 남부 지역에서 과소 모의하는 경향이 기계학습으로 후처리한 결과에서 개선되었다. 종합적으로 전국 10개 지점의 대기환경연구소에서 관측한 $PM_{2.5}$와 구성성분 농도와 비교한 결과, 기존 CMAQ 모델의 모의 결과보다 기계학습으로 후처리한 결과가 농도 분포를 현실적으로 개선하였다.

한편 기계학습의 후처리 결과는 해양에서의 $PM_{2.5}$ 농도가 높고 육지와 해양의 경계가 뚜렷하게 나타난다는 문제점이 있다. $PM_{2.5}$의 농도 분포를 보면 영동지역의 낮은 농도에 비해 동해안 연안 지역에서 높은 농도를 보인다. 이는 기계학습 모델의 입력자료로 사용되는 기상 변수들이 육지와 해양에서 변화폭이 크기 때문에 육지와 해양의 $PM_{2.5}$ 농도 간 차이가 나타난 것으로 분석된다.

기계학습으로 후처리한 $PM_{2.5}$ 구성성분의 농도를 합산한 결과가 이상적으로는 $PM_{2.5}$ 질량 농도와 같아야 한다. 그러나 본 연구에서는 각 구성성분을 개별적으로 측정 자료를 학습하여 모델을 구축하였기 때문에 후처리한 구성성분의 농도 합이 후처리한 $PM_{2.5}$ 질량 농도와 같지 않을 수 있다. 이를 확인하기 위해, 기계학습으로 후처리한 $PM_{2.5}$ 질량 농도와 각 구성성분(NO_3^-, SO_4^{2-}, NH_4^+, Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, OC, EC)을 개별적으로 후처리한 후 이를 합산한 $PM_{2.5}$ 농도($PM_{2.5}$ combine)를 비교하였다. 본 연구에서 $PM_{2.5}$ 성분 중 금속 성분, 토양 기원 성분 그리고 식별되지 않은 성분 농도들은 고려하지 않았다. 때문에, 기계학

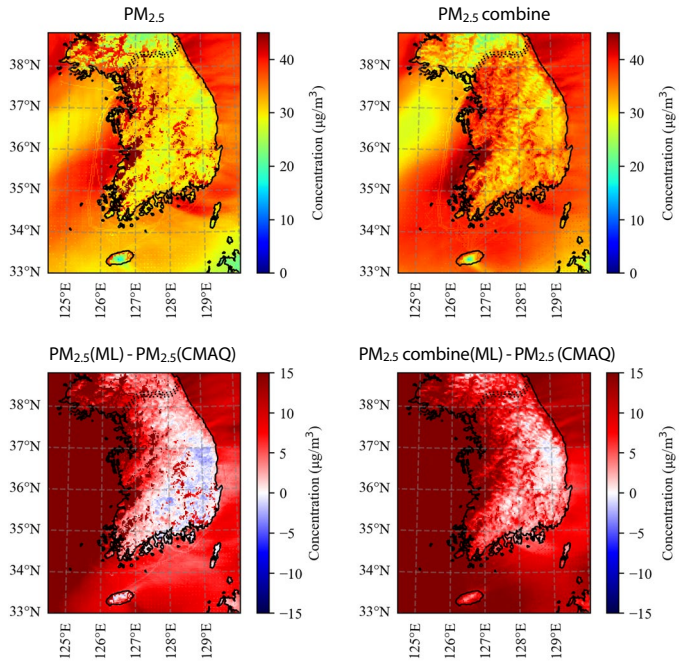

Fig. 9. The $PM_{2.5}$ concentration fields in the CMAQ-ML model at the surface level averaged from 10 KST 25 to 11 KST 26 February, 2022 (top). The concentration deviations obtained by subtracting CMAQ model data from the CMAQ-ML model data at the surface level averaged from 10 KST 25 to 11 KST 26 February, 2022 (bottom).

습으로 후처리 되지 않은 성분들을 반영하기 위해 기존 CMAQ 모델의 $PM_{2.5}$ 질량 농도와 성분 농도 합산 간 편차를 보정하였다.

그림 9는 $PM_{2.5}$ 질량 농도를 후처리한 결과, $PM_{2.5}$ 구성성분을 개별적으로 후처리한 후 합산한 결과, 그리고 이들과 기존 CMAQ 모델의 $PM_{2.5}$ 질량 농도 간 편차를 나타내었다. $PM_{2.5}$ 질량 농도를 후처리한 결과는 기존 CMAQ 모델의 $PM_{2.5}$ 질량 농도보다 서해안을 포함한 수도권, 중부권에서 높아지는 반면, 영남권에서는 낮아지는 경향을 모습을 보였다. 반면, 각 구성성분을 개별적으로 후처리한 후 이를 합산한 $PM_{2.5}$ 농도는 기존 CMAQ 모델의 $PM_{2.5}$ 질량 농도보다 전반적으로 높아지는 경향을 보였다. 이번 한 사례만으로 모든 기계학습의 특성을 일반화할 수는 없으나, $PM_{2.5}$ 질량 농도만을 학습하여 후처리하는 것에 비해 각 구성성분을 개별적으로 후처리하는 방법은 전국적으로 동일하게 적용하기보다 각 권역의 특성을 반영하기 위한 개선 노력이 더 필요함을 의미한다. 또한, 고농도 $PM_{2.5}$ 사례가 외부 유입 요인에 의한 기여가 많은 것을 감안할 때 우리나라의 지형적인 영향과 기상학적인 영향도 면밀하게 검토한 후 학습자료로 활용할 필요가 있다(Jeong *et al.*, 2024).

3.4 $PM_{2.5}$ 농도 후처리 방법 간 모의 성능 비교

$PM_{2.5}$ 및 성분 농도 후처리에 대한 검증은 전체 데이터 중 일부를 무작위로 추출하는 방법으로 수행했다. 이와 같은 방법은 지상 관측이 이뤄지지 않는 지역에서 모의 재현성을 평가하는 데 한계가 있다. 따라서 관측이 없는 지역에서도 모델 성능을 검증하기 위해 검증용 데이터를 별도로 구축하였다. 기계학습 모델 학습 시 연구 대상 전 기간을 대상으로 특정 지점을 학습자료에서 제외하고 모델을 구축하였으며, 제외한 지점을 검증용 자료로 활용하였다. 따라서 특정 지점에서 2022년 1월 1일부터 3월 31일까지의 자료가 검증에 사용되었다. 이를 통해 학습되지 않은 지점에서도 $PM_{2.5}$ 농도 후처리 성능이 일관되게 나타나는지 평가하고자 하였다. 또한 앞서 $PM_{2.5}$ 농도를 직접 후처리한 결과와 $PM_{2.5}$ 성분을 개별적으로 후처리한 후 합산한 결과 간 차이가 발생하였기 때문에 어느 방법이 기존 CMAQ 모델의 모의 결과를 더욱 개선하는지 비교해 볼 필요가 있다. 따라서 기존 CMAQ 모델의 $PM_{2.5}$ 농도, 기계학습(ML)으로 후처리한 $PM_{2.5}$ 농도, 기계학습으로 성분별 후처리한 후 합산 방식(combine_ML)으로 산출된 $PM_{2.5}$ 농도를 측정 농도와 비교하였다.

검증 대상 지역은 내륙 지역에서 고농도 현상이 빈번하게 발생하는 수도권과 호남권, 내륙 지역 외의 해상에 위치한 백령도로 선정하였다. 해당 지역은 타 권역에 비해 외부 유입으로 의한 고농도 미세먼지 발생 빈도가 높은 편이다(Yeo and Kim, 2019). 때문에 타 권역보다 고농도 사례에서 후처리 적용 전후에 농도 변화가 크게 나타난다. 같은 이유로 울산, 제주 등은 상대적으로 고농도 사례가 적기 때문에 성능 개선 여부를 평가하기에 적합하지 않다고 판단하였다.

수도권(서울)에서는 기계학습으로 $PM_{2.5}$ 농도를 후처리한 결과의 상관계수가 0.86, RMSE는 10.68 $\mu g/m^3$로 우수한 성능을 보였다. 반면, 호남권(광주)에서는 개별 성분을 후처리한 후 합산한 $PM_{2.5}$ combine 농도의 상관계수가 0.83, MAE가 7.28 $\mu g/m^3$, RMSE는 10.49 $\mu g/m^3$로 우수한 성능을 보였다. 백령도는 기계학습으로 $PM_{2.5}$ 농도를 후처리한 결과의 상관계수가 0.73,

Table 3. Model performance evaluation for $PM_{2.5}$ in Seoul, Gwangju, and Baengnyeong from 1 January to 31 March, 2022.

	r	MAE ($\mu g/m^3$)	RMSE ($\mu g/m^3$)
(a) Seoul			
CMAQ	0.72	12.97	18.09
ML	0.86	8.07	10.68
combine_ML	0.85	7.70	10.95
(b) Gwangju			
CMAQ	0.69	12.93	17.62
ML	0.80	8.07	10.98
combine_ML	0.83	7.28	10.49
(c) Baengnyeong			
CMAQ	0.68	9.38	14.37
ML	0.73	9.29	12.70
combine_ML	0.73	9.38	12.57

MAE가 9.29 μg/m^3, RMSE는 12.70 μg/m^{3}로 나타났으며, $PM_{2.5}$ combine 농도의 상관계수는 0.73, MAE가 9.38 μg/m^3, RMSE는 12.57 μg/m^{3}로 기계학습으로 $PM_{2.5}$ 농도를 후처리한 결과와 $PM_{2.5}$ combine 농도의 상관계수가 유사하게 나타났다(표 3). 내륙에 위치한 수도권과 호남권에서는 기존 CMAQ 모의 농도에 비해 기계학습을 적용하였을 때 검증지표가 개선되는 것을 확인할 수 있었다. 반면 백령도 지점은 기계학습 기법을 적용하였을 때 기존 CMAQ 모델 성능보다 크게 개선되지는 않았다.

학습에 사용되지 않은 지점을 대상으로 기계학습을 이용하여 후처리한 $PM_{2.5}$ 및 구성성분 농도를 검증한 결과에서 모두 기존 CMAQ 모델보다 우수한 성능을 보였으나, $PM_{2.5}$ 질량 농도를 후처리한 결과와 $PM_{2.5}$ 구성성분을 개별적으로 후처리한 후 합산한 결과 중 어느 방법이 더 우수한지는 지역에 따라 상이하게 나타났으며 지역의 특성을 면밀하게 반영할 필요가 있음을 다시 한번 확인하였다.

기계학습을 이용한 후처리 방법 간 미세먼지 농도 등급 모의 정확도를 비교하고자 $PM_{2.5}$ 농도 등급별 모의 정확도와 고농도 감지 확률(Probability of detection, POD)을 표 4에 비교하여 제시하였다. $PM_{2.5}$ 농도 등급은 에어코리아(AirKorea) 1시간 간격 $PM_{2.5}$ 농도 기준를 참고하여 좋음(0~15 μg/m^3), 보통(16~35 μg/m^3), 나쁨(36~75 μg/m^3), 매우나쁨(76 μg/m^3 이상)과 같이 4개의 등급으로 분류하였다. 수도권(서울)에서 기존 CMAQ 모델의 POD는 30%로 낮은 모의 성능을 나타냈으나, 기계학습 방법을 적용한 결과에서는 70%로 향상되었다. 기계학습으로 개별 구성성분을 후처리한 후 농도를 합산한 $PM_{2.5}$ combine 방식의 POD도 64%로, 기존 CMAQ 모델 결과 대비 우수한 성능을 보였다. 수도권(서울)의 $PM_{2.5}$ 등급이 매우나쁨(Serious)인 경우로 한정할 때, 기존 CMAQ 모델의 모의 정확도는 12.7%로 매우 낮은 반면, 기계학습으로 후처리한 방법의 모의 정확도는 65.8%, 개별 구성성분을 후처리한 후 농도를 합산한 $PM_{2.5}$ combine 방식의 모의 정확도는 48.1%로 나타났다. 호남권(광주)의 경우 기존 CMAQ 모델의 POD는 20%였으나, 기계학습으로 후처리한 결과의 POD는 70%로 향상되었다. 또한 $PM_{2.5}$ combine 방식의 POD도 60%로 기존 CMAQ 모델보다 모의 성능이 우수하였다. 호남권(광주)의 $PM_{2.5}$ 등급이 매우 나쁨인 경우로 한정할 때, 기존 CMAQ 모델의 모의 정확도는 0%로 나타났다. 이는 호남권에서 매우 나쁨 등급은 전체 중 약 2.4%로 사례가 적어 예측 정확도가 낮은 것으로 분석된다. 기계학습으로 후처리한 결과의 모의 정확도는 10.3%로 다소 향상되었으며, combine ML 방식의 모의 정확도도 2.6%를 보였다. 선행 연구에 따르면 기계학습을 통

Table 4. POD and simulation accuracy for $PM_{2.5}$ in Seoul, Gwangju, and Baengnyeong from 1 January to 31 March, 2022.

	POD (%)	Serious (%)	Bad (%)	Moderate (%)	Good (%)
(a) Seoul					
CMAQ	30	12.70	18.90	33.10	94.80
ML	70	65.80	61.60	79.20	41.30
combine_ML	64	48.10	55.60	80.60	55.90
(b) Gwangju					
CMAQ	20	0.00	18.30	26.80	97.40
ML	70	10.30	66.20	79.10	46.20
combine_ML	60	2.60	56.40	72.20	71.50
(c) Baengnyeong					
CMAQ	21	8.89	29.85	37.45	95.86
ML	35	0.00	60.20	67.45	55.50
combine_ML	33	0.00	56.22	71.57	48.64

한 미세먼지 예측 시 고농도 사례 자체가 적기 때문에 학습이 적어 예측 정확도가 낮은 것으로 알려져 있다 (Yu, 2024; Cho *et al.*, 2020). 백령도에서 기존 CMAQ 모델의 POD는 21%, 기계학습으로 후처리한 결과의 POD는 35%, $PM_{2.5}$ combine 방식의 POD는 33%로 내륙 지역에 비해 기계학습으로 인한 모의 정확도 개선 정도가 낮았다. 백령도의 $PM_{2.5}$ 등급이 매우 나쁨인 경우로 한정할 때, 기존 CMAQ 모델의 모의 정확도는 8.89%인 반면, 기계학습으로 후처리한 결과와 combine ML 방식의 모의 정확도는 0%로 고농도 모의 정확도가 낮아졌다.

호남권과 백령도의 경우 수도권보다는 여전히 낮은 모의 정확도임에도 기존 CMAQ 모델보다 기계학습 방식으로 후처리한 결과가 더 나은 성능을 보인다는 공통점을 찾을 수 있었다. 전반적으로 $PM_{2.5}$ 질량 농도를 직접 학습하여 개선하는 방식보다 개별 구성성분을 각각 학습하여 개선하는 방식이 아직은 효율적이지 못한 결과를 보인다. 이는 각 지역의 대기 중 물리화학 과정, 지형 및 기상의 요인 등을 충분히 반영하지 못해서일 것으로 판단되며 이를 개선하기 위한 노력이 앞으로도 더 필요하다. 또한 기계학습 방식으로 후처리한 결과는 관측이 없는 지역에서도 적용이 가능하나, 해상에서는 모의 정확도가 낮게 나타나 관측 지점이 많은 내륙 지역에서 적용성이 더 큰 것으로 판단된다.

4. 결 론

본 연구는 기계학습 기법을 통해 WRF-CMAQ 모델의 $PM_{2.5}$ 및 구성성분 농도 모의 성능을 개선하고자 하였다. 기계학습 알고리즘으로는 트리 기반의 RF, ET, XGBoost, LGBM을 사용하였으며, 그중 ET 알고리즘이 가장 우수한 성능을 보였다. 전국 10개 지점의 대기환경연구소 측정 자료와 비교한 결과에서 기계학습 모델을 이용하여 후처리한 결과는 기존 CMAQ 모델 결과에 비해 높은 정확성을 보였다. 특히 $PM_{2.5}$에서 상대적으로 높은 비중을 차지하는 NO_3^-, SO_4^{2-}, NH_4^+의 모의 성능이 상관계수 기준으로 기존 CMAQ 모델의 0.7 이하에서 기계학습 후처리 결과의 0.94 이상으로 크게 개선되었다. 고농도 $PM_{2.5}$ 사례 기간을 대상으로 분석한 결과, 기존 CMAQ 모델은 측정 결과와는 달리 주로 영남권에서 고농도 $PM_{2.5}$ 현상을 모의한 반면, 기계학습으로 후처리한 결과는 측정 결과와 유사하게 수도권, 경기권, 호남권에서의 고농도 $PM_{2.5}$ 현상을 나타냈다. 이와 함께 대기환경연구소가 부재한 지점을 가정하여 모의 성능을 검증하기 위해 수도권, 호남권, 백령도를 대상으로 기계학습 구축 시 각 지점 데이터를 학습에서 제외하고 모의 성능을 분석하였다. 학습에 사용되지 않는 지점에서도 기존 CMAQ 모델의 $PM_{2.5}$ 모의 성능보다 기계학습으로 후처리한 $PM_{2.5}$ 및 구성성분 모의 성능이 우수하게 나타났다. 특히 미세먼지 농도 등급이 나쁨 이상일 때 내륙 지역에서는 $PM_{2.5}$ 구성성분 농도를 합산한 방법보다 $PM_{2.5}$ 질량 농도를 후처리한 방법의 모의 성능이 높게 나타났다. 이는 기계학습을 통해 CMAQ 모델의 $PM_{2.5}$ 구성성분별 모의 성능을 향상시켰음에도 불구하고, 구성성분 간 상호작용을 고려하지 못하여 성분별로 합산하였을 때 예측 성능이 낮아진다는 한계점을 보여준다. 또한 해양에서는 나쁨 이상일 때 내륙 지역에 비해 기계학습 후보정 결과의 모의 성능이 크게 개선되지 않았다. 이는 아직 해상에서 기계학습을 통한 후보정 과정이 지형 및 기상 요인을 충분히 반영하지 못함을 의미한다. 본 연구는 내륙 지역에 한하여 기계학습 기법으로 CMAQ 모델의 $PM_{2.5}$ 구성성분 농도의 모의 성능을 개선하고 지상 관측이 없는 지점에서도 신뢰할 수 있는 결과를 제공할 수 있음을 확인하였다. 이는 미세먼지의 발생원과 대기 중 2차 생성 과정을 규명하는 데 대기환경연구소 측정 자료로 개선한 CMAQ 모델 결과가 효과적으로 활용될 수 있음을 의미한다. 따라서 이러한 방법을 좀 더 개선 및 보완한다면 향후 대기질 관리 정책을 수립하는 데 활용될 수 있을 것으로 기대된다.

감사의 글

이 연구는 국립환경과학원에서 주최한 제3회 대학(원)생 미세먼지연구아이디어공모전으로 수행되었습니다(NIER-2024-03-00-005). 이 성과는 정부(환경부)의 재원으로 한국환경산업기술원의 미세먼지관리특성화대학원 사업의 지원 및 환경부의 재원으로 국립환경과학원의 지원(NIER-2021-03-03-007)을 받아 수행한 연구입니다. 배출량 자료를 제공해주신 국가미세먼지정보센터에 감사드립니다.

References

Appel, K.W., Bash, J.O., Fahey, K.M., Foley, K.M., Gilliam, R.C., Hogrefe, C., Hutzell, W.T., Kang, D., Mathur, R., Murphy, B.N., Napelenok, S.L., Nolte, C.G., Pleim, J.E., Pouliot, G.A., Pye, H.O.T., Ran, L., Roselle, S.J., Sarwar, G., Schwede, D.B., Sidi, F.I., Spero, T.L., Wong, D.C. (2021) The Community Multiscale Air Quality (CMAQ) model versions 5.3 and 5.3.1: System updates and evaluation, Geoscientific Model Development, 14(5), 2867-2897. https://doi.org/10.5194/gmd-14-2867-2021

Bae, H.-J., Lee, S.M., Jung, D.-W., Oh, G.-L., Kim, S.J., Lee, J.-T. (2020) Study on the Health Effects of $PM_{2.5}$ Constituents for Health Risk Reduction Management Plan. Korea Environment Institute Environment Forum, 24(4), 1-19. https://www.kei.re.kr/elibList.es?mid=a10102010000&elibName=environmentalforum&class_id=&act=view&c_id=727263&rn=10&nPage=1&keyField=&keyWord=

Bae, M., Woo, J.-H., Kim, S. (2023) Seasonal PM Management: (II) How Low $PM_{2.5}$ Concentration in South Korea can be Achieved?, Journal of Korean Society for Atmospheric Environment, 39(1), 9-23, (in Korean with English abstract). https://doi.org/10.5572/kosae.2023.39.1.9

Benjey, W., Houyoux, M., Susick, J. (2001) Implementation of the SMOKE Emission Data Processor and SMOKE Tool Input Data Processor in Models-3. Presented at the Emission Inventory Conference, Denver, CO, USA, 1-4 May 2001; p. 15.

Carter, W.P.L. (2010) Development of the SAPRC-07 chemical mechanism, Atmospheric Environment, 44(40), 5324-5335. https://doi.org/10.1016/j.atmosenv.2010.01.026

Cha, Y., Song, C.K., Jeon, K.H., Yi, S.M. (2023) Factors affecting recent $PM_{2.5}$ concentrations in China and South Korea from 2016 to 2020, Science of The Total Environment, 881, 163524. https://doi.org/10.1016/j.scitotenv.2023.163524

Chen, F., Dudhia, J. (2001) Coupling an advanced land surface - hydrology model with the Penn State - NCAR MM5 modeling system. Part I: Model implementation and sensitivity, Monthly Weather Review, 129(4), 569-585.

Cho, K.W., Jung, Y.J., Lee, J.S., Oh, C.H. (2020) Separation prediction model by concentration based on deep neural network for improving PM_{10} forecast accuracy, Journal of the Korea Institute of Information and Communication Engineering, 24(1), 8-14, (in Korean with English abstract). https://doi.org/10.6109/jkiice.2020.24.1.8

Choi, M.W., Lee, J.H., Woo, J.W., Kim, C.H., Lee, S.H. (2019) Comparison of $PM_{2.5}$ chemical components over East Asia simulated by the WRF-Chem and WRF/CMAQ models: On the models' prediction inconsistency, Atmosphere, 10(10), 618.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M.B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L.J., Schwartz, J. (2019) An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution, Environment International, 130, 104909. https://doi.org/10.1016/j.envint.2019.104909

Do, W.G., Kim, D.Y., Song, H.J., Cho, G.J. (2023) A Study on the $PM_{2.5}$ forecasting Method in Busan Using Deep Neural Network, Journal of Environmental Science International, 32(8), 595-611, (in Korean with English abstract).

Geurts, P., Ernst, D., Wehenkel, L. (2006) Extremely randomized trees, Machine Learning, 63(1), 3-42. https://doi.org/10.1007/s10994-006-6226-1

Ghahremanloo, M., Choi, Y., Sayeed, A., Salman, A.K., Pan, S., Amani, M. (2021) Estimating daily high-resolution $PM_{2.5}$ concentrations over Texas: Machine Learning approach, Atmospheric Environment, 247. https://doi.org/10.1016/j.atmosenv.2021.118209

Guenther, A., Karl, T., Harley, P., Wiedinmyer, C., Palmer, P.I., Geron, C. (2006) Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature), Atmospheric Chemistry and Physics, 6(11), 3181-3210. https://doi.org/10.5194/acp-6-3181-2006

Guyon, I., Elisseeff, A. (2003) An introduction to variable and feature selection, Journal of Machine Learning Research,

3, 1157-1182.

Hong, S.Y., Noh, Y., Dudhia, J. (2006) A new vertical diffusion package with an explicit treatment of entrainment processes, Monthly Weather Review, 134(9), 2318-2341.

Huang, C., Niu, T., Wu, H., Qu, Y., Wang, T., Li, M., Li, R., Liu, H. (2023) A Data Assimilation Method Combined with Machine Learning and Its Application to Anthropogenic Emission Adjustment in CMAQ, Remote Sensing, 15(6), 1711. https://doi.org/10.3390/rs15061711

Jeong, H.-Y., Kwak, K.-H. (2022) Review on the Prediction Studies of $PM_{2.5}$ Using Machine Learning, Journal of the Environment, 15(1), 41-48, (in Korean with English abstract).

Jeong, S., Kang, Y.-H., Kim, E., Kim, S. (2024) Terrain Effect on Provincial $PM_{2.5}$ Concentrations in South Korea, Journal of Korean Society for Atmospheric Environment, 40(6), 628-645, (in Korean with English abstract). https://doi.org/10.5572/kosae.2024.40.6.628

Kang, Y.-H., Kim, E., You, S., Bae, M., Son, K., Kim, B.-U., Kim, H.-C., Kim, S. (2021) Source Sectoral Impacts on Provincial $PM_{2.5}$ Concentrations based on the CAPSS 2016 using the CMAQ Model, Journal of Korean Society for Atmospheric Environment, 37(1), 17-44, (in Korean with English abstract). https://doi.org/10.5572/kosae.2021.37.1.017

Kim, B.-Y., Lim, Y.-K., Cha, J.W. (2022) Short-term prediction of particulate matter (PM_{10} and $PM_{2.5}$) in Seoul, South Korea using tree-based machine learning algorithms. Atmospheric Pollution Research, 13(10). https://doi.org/10.1016/j.apr.2022.101547

Kim, Y.-I., Lee, K.-H. (2023). Accuracy Analysis of Machine Learning Methods for Predicting PM Concentration, Journal of Korean Society for Atmospheric Environment, 39(2), 149-164, (in Korean with English abstract). https://doi.org/10.5572/kosae.2023.39.2.149

Kim, Y.-I., Lee, K.-H., Park, S.-H. (2023) Application and Evaluation of Machine Learning Techniques for Real-time Short-term Prediction of Air Pollutants, Journal of Korean Society for Atmospheric Environment, 39(1), 107-127, (in Korean with English abstract). https://doi.org/10.5572/kosae.2023.39.1.107

Korea Meteorological Administration (KMA) (2022) Climate Analysis Report: February 2022, Climate Analysis Report, 2022(2), 1-13. https://www.kma.go.kr/download_02/ellinonewsletter_2022_02.pdf

Kwak, K.-H. (2022) No. 26 Meteorological factors of particulate matter. Particulate matter insight. December 2022. 3-4.

Lee, K.-B., Kim, S.-D., Kim, D.-S. (2015) Ion Compositional Existence Forms of PM_{10} in Seoul Area, Journal of Korean Society of Environmental Engineers, 37(4), 197-203, (in Korean with English abstract). https://doi.org/10.4491/ksee.2015.37.4.197

Liu, S., Geng, G., Xiao, Q., Zheng, Y., Liu, X., Cheng, J., Zhang, Q. (2022) Tracking Daily Concentrations of $PM_{2.5}$ Chemical Composition in China since 2000, Environmental Science & Technology, 56(22), 16517-16527. https://doi.org/10.1021/acs.est.2c06510

Louis, J.F. (1979) A parametric model of vertical eddy fluxes in the atmosphere, Boundary-Layer Meteorology, 17(2), 187-202.

Ministry of Environment (MOE) (2022) Fourth seasonal fine dust management plan implementation plan. https://www.me.go.kr/home/web/policy_data/read.do?pagerOffset=0&maxPageItems=10&maxIndexPages=10&searchKey=title&searchValue=%EA%B3%84%EC%A0%88%EA%B4%80%EB%A6%AC%EC%A0%9C&menuId=10262&orgCd=&condition.toInpYmd=null&condition.fromInpYmd=null&condition.deleteYn=N&condition.deptNm=null&seq=7992

National Air Emission Inventory and Research Center (NAIR) (2023) 2021 National Air Pollutant Emissions Inventory. https://www.air.go.kr

National Institute of Environmental Research (NIER) (2022) 2021 Annual Report of Air Quality Research Center. https://www.air.go.kr

Nava, S., Calzolai, G., Chiari, M., Giannoni, M., Giardi, F., Becagli, S., Severi, M., Traversi, R., Lucarelli, F. (2020) Source Apportionment of $PM_{2.5}$ in Florence (Italy) by PMF Analysis of Aerosol Composition Records, Atmosphere, 11(5), 484. https://doi.org/10.3390/atmos11050484

Park, J.S., Kim, C.H., Lee, J.J., Kim, J.H., Hwang, U.H., Kim, S.D. (2010) A study on the chemical mass composition of particle matter in Seoul, Journal of Korean Society of Urban Environment, 10(3), 293-303, (in Korean with English abstract).

Park, S., Kim, M., Im, J. (2021) Estimation of ground-level PM_{10} and $PM_{2.5}$ concentrations using boosting-based machine learning from satellite and numerical weather prediction data, Korean Journal of Remote Sensing, 37(2), 321-335, (in Korean with English abstract). https://doi.org/10.7780/kjrs.2021.37.2.11

Pleim, J.E. (2007) A Combined Local and Nonlocal Closure Model for the Atmospheric Boundary Layer. Part I: Model Description and Testing, Journal of Applied Meteorology and Climatology, 46(9), 1383-1395. https://doi.org/10.1175/jam2539.1

Smyth, S.C., Jiang, W., Yin, D., Roth, H., Giroux, É. (2006) Evalua-

tion of CMAQ O_3 and $PM_{2.5}$ performance using Pacific 2001 measurement data, Atmospheric Environment, 40(15), 2735-2749. https://doi.org/10.1016/j.atmosenv.2005.10.068

Thongthammachart, T., Araki, S., Shimadera, H., Eto, S., Matsuo, T., Kondo, A. (2021) An integrated model combining random forests and WRF/CMAQ model for high accuracy spatiotemporal $PM_{2.5}$ predictions in the Kansai region of Japan, Atmospheric Environment, 262. https://doi.org/10.1016/j.atmosenv.2021.118620

Wu, Q., Tang, X., Kong, L., Liu, Z., Chen, D., Lu, M., Wu, H., Shen, J., Wu, L., Pan, X., Li, J., Zhu, J., Wang, Z. (2021) Model Evaluation and Uncertainty Analysis of $PM_{2.5}$ Components over Pearl River Delta Region Using Monte Carlo Simulations, Aerosol and Air Quality Research, 21(1), 200075. https://doi.org/10.4209/aaqr.2020.02.0075

Xiong, K., Xie, X., Huang, L., Hu, J. (2024) Improved O_3 predictions in China by combining chemical transport model and multi-source data with machining learning techniques, Atmospheric Environment, 318. https://doi.org/10.1016/j.atmosenv.2023.120269

Yamartino, R.J. (1993) Nonnegative, conserved scalar transport using grid-cell-centered, spectrally constrained Blackman cubics for applications on a variable-thickness mesh, Monthly Weather Review, 121(3), 753-763.

Yeo, M.J., Kim, Y.P. (2019) Trends of the PM_{10} Concentrations and High PM_{10} Concentration Cases in Korea, Journal of Korean Society for Atmospheric Environment, 35(2), 249-264, (in Korean with English abstract). https://doi.org/10.5572/kosae.2019.35.2.249

Yu, S. (2024) Analysis of Input Factor Contributions Characteristics by Index in $PM_{2.5}$ Forecast Using Layer-wise Relevance Propagation and DNN, Journal of Korea Multimedia Society, 27(5), 585-599, (in Korean with English abstract). https://doi.org/10.9717/kmms.2024.27.5.585

Zhang, Z., Xu, B., Xu, W., Wang, F., Gao, J., Li, Y., Li, M., Feng, Y., Shi, G. (2022) Machine learning combined with the PMF model reveal the synergistic effects of sources and meteorological factors on $PM_{2.5}$ pollution, Environmental Research, 212. https://doi.org/10.1016/j.envres.2022.113322

Authors Information

김지민(강원대학교 환경학과 박사과정)
(jimin1118@kangwon.ac.kr)

최민서(강원대학교 미세먼지통합관리학과 석사과정)
(rane127@kangwon.ac.kr)

전예지(강원대학교 환경학과 석사과정)
(202110080@kangwon.ac.kr)1),

김태희(강원대학교 미세먼지통합관리학과 박사과정)
(spartan090@kangwon.ac.kr)

곽경환(강원대학교 환경학과 · 미세먼지통합관리학과 부교수)
(khkwak@kangwon.ac.kr)

이그림(국립환경과학원 대기환경연구부 대기환경연구과 연구사)
(greemlee@korea.kr)

강병철(국립환경과학원 대기환경연구부 대기환경연구과 연구원)
(bcblue66@korea.kr)

정선아(국립환경과학원 대기환경연구부 대기환경연구과 연구원)
(1004sun04@korea.kr)