수치모델과 지리정보를 활용한 부산지역 고해상도 (초)미세먼지, 오존 자료 생성 방법 연구

도우곤

대기환경연구부 대기진단평가팀

A Study on How to Generate High Resolution O_3, PM10, and PM2.5 Data using Numerical Model and GIS in Busan

Do Woo-gon

Air Quality Monitoring and Assessment Team

Abstract

The purpose of this study is to improve the hourly prediction results of O_3, PM10, and PM2.5 from the air quality diagnosis and evaluation system operated by the Busan Institute of Health and Environment in real time. The air quality diagnosis and evaluation system is based on the photochemical numerical model, CMAQ and includes a 3-day forecast at the end of the model's calculation. The photochemical numerical model basically includes uncertainty due to uncertainty of input data and simplification of physical and chemical processes. To overcome this uncertainty, this study applied SVM, a machine learning technique, to the results of the numerical model. As a result of applying SVM, the R^2 of the model was significantly improved compared to before application, with O_3 from 0.30 to 0.69, PM2.5 from 0.27 to 0.65 and PM10 from 0.16 to 0.55. RMSE, which means the model error rate, was also significantly improved with O_3 0.010 ppm, PM10 12.9ug/m^3, and PM2.5 7.5ug/m^3 compared to 0.017, 22, and 14 before application.

Key words *: CMAQ, Machine learning, Model performance*

1. 서론

O_3, PM2.5 등 대기오염으로 인한 건강 위해성이 알려지면서 대기오염에 의한 건강상의 피해를 예방하고 이들 오염물질의 발생메카니즘과 공간적인 분포 등을 파악하기 위하여 광화학 수치모델을 활용한 연구가 활발하게 진행되고 있다. 최근에는 전산기술의 발달로 광화학 수치모델의 결과를 실시간으로 공개하거나 이를 활용하여 대기오염도를 사전에 예측하는 경우가 많은데 국외의 경우 미국의 국립해양대기청(NOAA)에서 The Community Multiscale Air Quality Modeling System(CMAQ) 모델을 기반으로 한 대기질 예보시스템(NAQFC)을 개발하여 지표면 먼지(dust)와 O_3의 예측농도를 National Weather Service[1]를 통하여 공개하고 있으며, 캐나다 환경청에서는 GEM(Global Environmental Multi-scale Modelling) 기상모델과 캐나다와 미국의 국가 배출목록을 입력자료로 하여 Air quality and CHemistry(MACH) 대기확산 모델을 활용한 O_3, PM2.5, PM10 예측농도를 Regional Air Quality Deterministic Prediction System (RAQDPS)[2]을 통하여 공개하고 있다. 영국의 경우는 AQUM이라는 모델링 시스템을 이용하여 최대 5일 후까지의 대기오염지수 분포를 홈페이지를 통하여 공개하고 있으며 인도의 지구과학부와 기상청에서는 System of Air Quality and Weather Forecasting And Research 모델링 시스템을 이용하여 인도 내의 주요 도시에 대한 내일의 대기오염도 지수를 공개하고 있다[3, 4]. 국내의 경우 국립환경과학원의 대기질통합예보센터에서 PM10, PM2.5 및 O_3 농도 등급의 예보를 2014년부터 시행해오고 있으며 오염물질의 농도 등급과 더불어 한반도 전체 권역에 대한 CMAQ 모델링 결과를 시각화하여 에어코리아를 통하여 공개하고 있다[5]. 한편 지자체 중 최초로 부산광역시에서는 CMAQ 모델을 기반으로 하는 대기질 진단평가시스템을 구축하여 2017년부터 지역내의 대기오염물질 농도를 예측하는데 활용하고 있다[6, 7]. 광화학 수치모델을 이용하여 원하는 지역의 대기오염물질을 예측하기 위해서는 우선 기상모델을 이용하여 모의하고자 하는 영역의 3차원 기상장을 계산하고 모델에 필요한 시간별 및 종별 배출량 자료를 생성해야 한다. 이렇게 확보된 기상장과 각 화학종별 배출량 자료를 광화학 수치모델의 입력자료로 사용하여 최종적으로 시간별 대기질 농도를 예측할 수 있다. 이 과정에서 입력자료로 사용되는 배출량이나 초기 기상자료가 가지는 불확실성에 따라 기본적으로 모델링 결과에 오류가 포함되게 된다[8, 9]. 또한 모델에서 사용되는 화확반응 메카니즘 등 물리, 화학 작용의 수식화와 수치해석 방법 등에서도 오차가 포함되게 된다[10]. 다시말하면 광화학 수치모델의 결과를 직접적으로 대기오염 예측자료로 사용하기에는 실측되는 대기오염물질의 농도와 차이가 발생하기 때문에 다소 무리가 있으며 이는 배출량 입력자료의 불확실성, 초기 기상장과 기상 모사에 따른 불확실성, 모델에서 사용되는 대기오염물질 농도 초기장의 실제와의 차이 그리고 모델에서 사용되는 물리, 화학과정의 단순화 등 기본적인 문제에서 기인한다[11]. 따라서 이러한 광화학 수치모델의 기본적인 약점을 보완하기 위하여 입력자료 및 초기조건의 현실화, 모델 내 물리, 화학과정의 개선 등 다양한 시도가 지속적으로 이루어져 오고 있다[8, 9, 10, 11]. 이와 더불어 최근에는 대기오염물질의 시, 공간적인 예측에 머신러닝 기법을 활용하는 사례가 늘고 있다[12]. 대기오염물질 농도 분석을 위한 전통적인 통계모델은 활용 가능한 변수의 종류나 통계적인 특성에 많은 제약이 있었으나 인공신경망이나 서포트벡터머신 또는 랜덤포레스트 같은 머신러닝 기법들은 변수의 비선형성에 포함된 불확실성을 극복하여 매우 높은 정확성을 보여주는 것으로 나타나고 있다[13]. Dutta and Jinsart(2021)은 인도 구와하티 지역의 PM10 농도를 다중선형회귀(MLR), 다중신경망(MLP), 분류.회귀 의사결정나무(CART)의 세 가지 방법으로 예측하였으며, Shahriar et al.(2020)은 방글라데시 주요 도시의 PM10과 PM2.5 농도를 대기오염물질 농도와 기상요소를 결합하여 다양한 머신러닝기법을 사용하여 예측하고 정확성을 비교하였다[14, 15]. Madhavi et al.(2014)은 뉴질랜드 오클랜드 지역에서 실시간 측정되는 기상요소를 인공신경망 모델에 적용하여 NO_2 농도를 예측한 바 있으며 Goulier et al.(2020)은 독일 Münster 지역의 NO, NO_2, NO_x와 O_3 농도를 실시간 측정자료와 교통량, 시계열 정보를 이용하여 인공신경망으로 예측을 하였다[16, 17]. 이와 더불어 국내에서도 머신러닝을 대기오염 예측에 활용한 다수의 사례가 있다[18, 19, 20]. 이들 연구결과들을 보면 예측된 대기오염변수들은 실측치와 상당한 일치도를 보이는 것으로 나타나 머신러닝이 광화학 수치모델의 약점을 보완할 수 있는 대안이 될 수 있다고 판단된다. 또한 최근에 Sayeed et al.(2021)은 CMAQ 모델의 결과에 인공신경망을 적용하여 우리나

라 255개 대기오염측정소의 시간별 오존농도를 성공적으로 예측하여 광화학 수치모델의 결과를 개선하는 방법을 제시하였다[21]. 본 연구는 부산광역시 보건환경연구원에서 실시간으로 운영중인 CMAQ 모델링 시스템의 O_3, PM10, PM2.5의 시간별 예측결과를 개선하는 것을 목적으로 한다. Sayeed et al.(2021)의 연구에서 제시된 것처럼 수치모델의 결과에 머신러닝기법을 적용하여 모델결과와 실측치와의 일치도를 높이고 산정된 관계식을 전체 모델영역에 적용하여 실측자료로 보정된 격자별 O_3, PM10, PM2.5 시간별 농도를 생성하여 대기오염측정소가 없는 지역의 대기질 농도를 상시 계산할 수 있도록 하였다. 이를 바탕으로 고해상도의 O_3, PM10, PM2.5 시간별 농도를 대기오염측정소의 추가 없이 생성하여 대기환경개선을 위한 기초자료로 제공하고자 한다.

2. 재료 및 방법

2.1. 진단평가시스템

대기질 진단평가시스템은 CMAQ 모델을 활용하여 실시간 모델링이 수행되는 시스템으로 부산광역시 보건환경연구원에서 2017년에 최초 도입하였다[7]. 모델링 영역은 동아시아 27km 격자 영역에서 한반도 9km 격자 영역, 영남권 3km 격자 영역, 최종적으로 부산권의 1km 격자 영역으로 4단계의 nesting 도메인으로 구성된다(Table 1, Fig. 1).

진단평가시스템의 기상입력자료는 NCEP의 Global Forecast System(GFS)와 기상청 국가기상 슈퍼컴퓨터 센터의 지역예보모델에서 생성된 Unified Model(UM) 자료를 Weather Research Forecast(WRF) 모델에 입력하여 두 가지 경우로 나누어 생성되고 있다 [23]. GFS는 NCEP에서 운영하고 있는 전구 기상예보 수치모델이며 하루에 4번 6시간 간격으로 실행되고 384시간의 예보자료를 공개하고 있다[22]. 기상청 국가기상 슈퍼컴퓨터 센터에서 생성된 수치모델 자료는 영국 통합모델(UM)을 기반으로 전지구예보모델, 지역예보모델, 국지예보모델로 구분된다. 진단평가시스템에서 사용되는 UM은 지역예보모델 자료를 사용하고 있으며 공간해상도는 수평으로 12km, 연직으로 약 80km까지 70층으로 구성되며, 3시간 간격으로 전지구예보모델로부터 경계장을 제공받아 1일 4회(00, 06, 12, 18UTC) 87시간 예측을 수행한다[23]. WRF 모델은 중규모 기상모델인 PSU/NCAR-MM5를 대체하기 위해 미국 대기과학연구소 National Center of Atmospheric Research(NCAR)와 National Center for Environmental Prediction(NCEP)이 공동으로 개발하였으며 2005년부터 미국 NOAA 산하 기관인 NCEP의 현업 모델로 사용되고 있고 세계적으로 널리 보급되어 많은 연구에 활용되고 있는 기상모델이다[24]. 광화학 모델링을 위한 배출량 자료를 생성하기 위해서 미국 EPA에서 제공하는 Sparse Matrix Operator Kernel Emissions(SMOKE) 모델을 적용하였다. SMOKE는 미국의 Environmental Modeling Center(EMC)에서 개발된 것으로 모델링에 필요한 배출량을 고효율로 계산할 수 있도록 배출량을 Matrix 구조체로 생성하는 배출량 모델링 시스템이다[25]. SMOKE 모델에 사용하는 동아시아 지역의 배출량자료로 중국의 Multi-resolution emission inventory for China(MEIC) 배출량, 중국을 제외한 아시아 지역은 Regional Emission inventory in ASia(REAS) 배출량 자료를 이용하였다. MEIC 배출량은 중국 Tsinghua University에서 개발되었으며 중국지역을 대상으로 수평해상도 0.25도로 power, industry, residential, transportation, agriculture의 5개 부문에 대하여 배출량을 공개하고 있다[26]. REAS 배출량은 Ohara et al.(2007)에 의해 최초로 개발되었으며 SO_2, NOx, CO, NMVOC 등 총 10개 물질에 대하여 아시아지역 전체에 대하여 0.25도 해상도로 배출량을 공개하고 있다[27, 28]. 국내배출량 자료는 환경부 국가미세먼지정보센터의 대기정책지원시스템(Clean Air Policy Support System, CAPSS)에서 공개되는 2012년 배출원별, 1km 격자별 배출량을 사용하였다[39]. 한편 산림, 초지 등 식생 배출원의 영향은 Biogenic Emission Inventory System version 3(BEIS3)과 Model of Emissions of Gases and Aerosols from Nature(MEGAN)을 적용하여 계산하였다[29, 30]. 시, 공간적인 대기오염물질의 농도변화를 계산하기 위해서 생성된 기상 및 배출량 입력자료는 광화학 수치모델인 CMAQ에 입력된다. CMAQ은 미국 EPA가 정한 규제모형 중에서 가장 많이 이용되는 3차원 광화학 오일러리안 대기질 모델이며, 대기 중 오염농도, 건성침적, 습성침적 등 여러 가지 물리적 과정과 대기 중에서 발생하는 광화학 반응 등 상세한 물리·화학 반응 모듈을 포함하고 있다[31, 32]. CMAQ의 큰 장점은 첫 번째로 모듈구조로 되어있다는 특징을 가지고 있어 각 서브 프로그램간 그리고 각 전처리 단계간의

상호 호환이 쉽고 효율적이다. 두 번째로 모델링 영역의 규모가 다양하여 국지규모에서 지역규모 모델링까지 다양하게 동시에 모델링이 가능하다. 세 번째 특성으로는 황화합물이나 오존화합물 뿐 아니라 최근 기후적 측면과 국지오염 측면에서 중요한 관심사가 되고 있는 2차생성 에어로졸까지 동시에 여러 오염물질을 모사할 수 있다[31, 32]. 대기질 진단평가시스템의 구조와 모델링 옵션은 Fig. 2, Table 2에 제시하였다.

Table 1. Grid configurations of modelling system

	Model domain	Grids information
Horizontal grids	Domain1	174 × 128 × 27km
	Domain2	67 × 82 × 9km
	Domain3	83 × 83 × 3km
	Domain4	78 × 70 × 1km
Vertical levels	sigma levels	1.000, 0.995, 0.990, 0.985, 0.970, 0.950, 0.930, 0.910, 0.880, 0.840, 0.800, 0.740, 0.700, 0.600, 0.450, 0.000

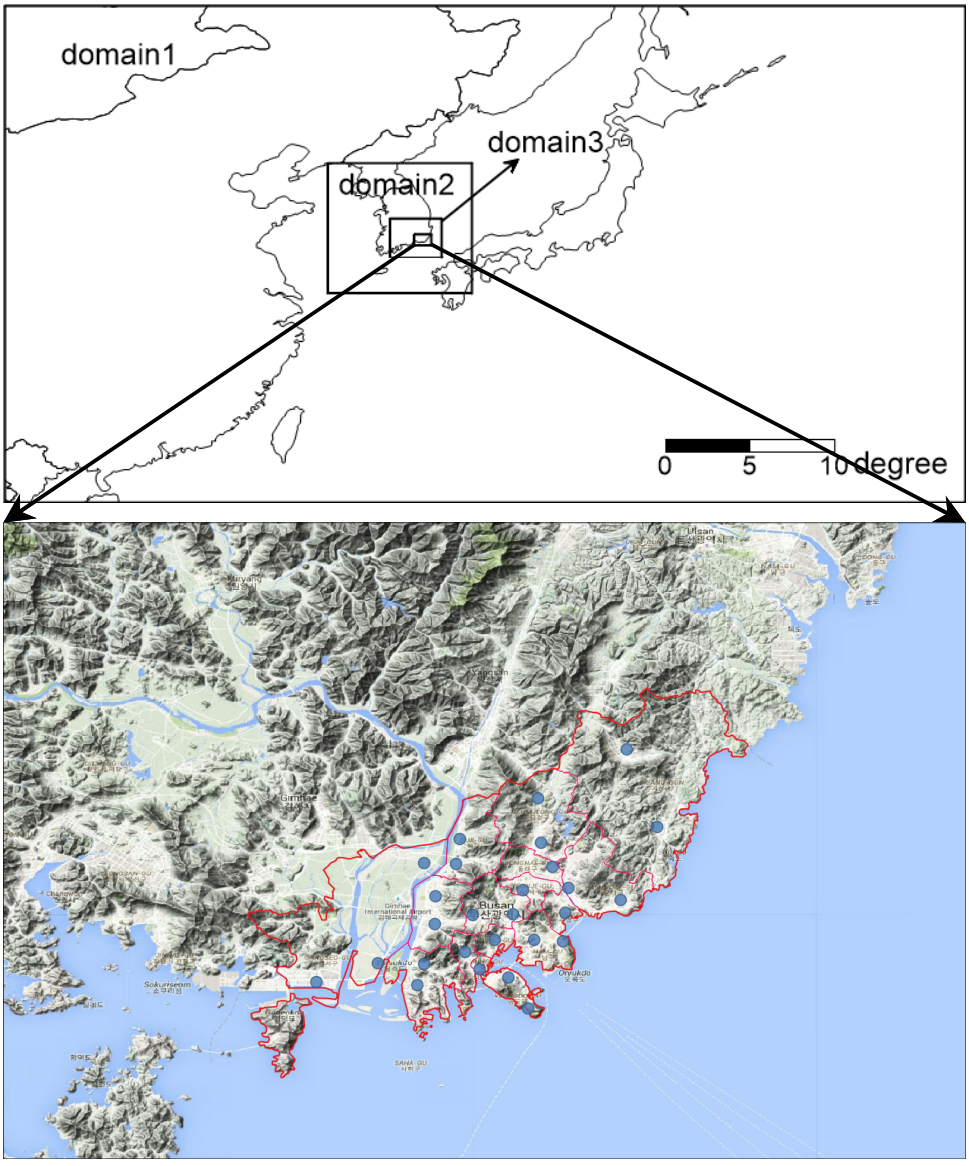

Fig. 1. Modeling domain for air quality diagnosis and evaluation system. Upper part shows locations of domain1 to domain3, lower part shows geographical features of domain4 and locations of air quality monitoring stations.

Table 2. Descriptions of model configurations

WRF physics options
• Microphysics option : WSM 6-calss graupel • Long wave radiation : RRTM • Short wave radiation : Goddard • Surface layer scheme : MM5 similarity • Land surface scheme : Noah Land Surface Model • PBL scheme : YSU • Cumulus parameterization : Kain-Fritsch
CMAQ options
• Horizontal advection : YAMO • Vertical advection : WRF • Horizontal diffusion : Multi-scale • Vertical diffusion : Eddy • Gas-phase chemistry : CB5 • Aerosol chemistry : AE5 • Dry deposition : M3Dry

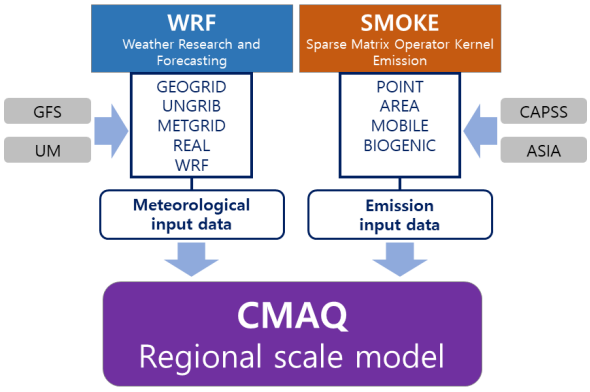

Fig. 2. Schematic diagram of the procedure in air quality diagnosis and evaluation system.

2.2. 대기오염측정자료

진단평가시스템의 O_3, PM10, PM2.5의 계산결과를 검증하고 모델결과의 개선을 위하여 부산광역시 보건환경연구원에서 운영 중인 도시대기측정소의 시간별 측정자료를 활용하였다[33]. 2019년에서 2020년까지의 지점별 시간자료를 모델링 결과와 비교하여 모델의 정확성을 검증하고 모델결과를 개선하기 위하기 위한 머신러닝의 입력자료로 활용하였다. 다음으로 2021년 시간별 자료를 활용하여 구축된 머신러닝 모델의 예측성을 평가하는데 활용하였다. 대기오염측정소는 2020년까지 구축된 총 27개의 도시대기측정소를 대상으로 하였으며 분석기간 중 신설되는 측정소의 자료도 분석에 포함하였다. 대기오염측정소의 위치는 Fig. 1과 같으며 각 항목별 측정방법은 대기오염공정시험기준을 따르며 주 1회 이상 정도관리를 수행하고 있다.

2.3. Support Vector Machine(SVM)

SVM은 Vapnik(1995)이 제안한 머신러닝 기법으로, 경험적 위험 최소화 원칙을 기반으로 하는 다른 통상적인 기계학습 기법과는 달리 구조적 위험 최소화를 기반으로 하여 일반화 오류의 상한을 최소화할 수 있는 머신러닝(machine learning) 기법이다[34]. 이 중에서 SVM은 Artificial neural network(ANN) 기법의 문제점으로 지적되는 과적합(overfitting) 문제를 벌칙(penalty)항을 이용하여 피할 수 있으며, 또한 함수 근사에 있어서 이상데이터(outlier)에 둔감하기 때문에 높은 일반화 성능을 가진다[34]. 따라서, 만약 동일한 데이터를 활용할 경우, 데이터의 특성에 따라 ANN에 비해 상대적으로 예측력이 우수한 모델의 구현이 가능한 장점이 있다[35]. SVM의 실행을 위해서 본 연구에서는 R 프로그램의 e1071 패키지를 사용하였으며 SVM에서 선택할 수 있는 4개의 커널중 비교적 높은 정확도를 나타내는 것으로 알려진 Radial Basis Function(RBF) 커널을 선택하였다[36]. RBF 커널의 파라미터들은 학습 오류의 최소화와 모델의 복잡성 사이의 값을 나타내는 Cost(C)와 일부 고차원 특성 공간으로의 비선형 매핑을 정의하는 gamma가 있는데 본 연구에서는 반복학습을 통하여 최적의 파라메타 값을 산정하여 모델에 적용하였다[37].

2.4. 모델의 적합성 평가 및 분석방법

진단평가시스템의 CMAQ 모델의 결과와 SVM으로 개선된 모델결과의 적합성은 관측자료와의 비교를 통하여 수행하였다. 모델치와 관측치의 적합성을 확인하는 변수들은 매우 다양하며 본 연구에서는 Mean Bias(MB), Root Mean Square Error(RMSE) 그리고 R^2 값을 계산하여 모델의 적합성을 평가하였다[38]. MB는 모델값과 측정값의 차이를 전 기간에 대하여 평균한 것으로 +값이면 측정값에 비해 모델값이 과대평가, -값이면 과소평가하는 것으로 판단할 수 있으며 RMSE는 모델값과 측정값의 평균 제곱근 오차로 측정값에 비하여 모델값이 어느 정도의 오차를 가지는지를 판단할 수 있다. R^2 값은 결정계수라고도 하며 모델값이 측정값에 어느 정도의 설명력을 가지는지를 의미하는 변수이다.

$$MB=\frac{1}{n}\sum_{i=1}^{n}(model(i)-Obs(i)) \quad (1)$$

$$RMSE=[\frac{1}{n}\sum_{i=1}^{n}(model(i)-Obs(i))^2]^{\frac{1}{2}} \quad (2)$$

$$R^2=[\frac{\sum_{i=1}^{n}(model(i)-\overline{model})(Obs(i)-\overline{Obs})}{\sqrt{\sum_{i=1}^{n}(model(i)-\overline{model})^2\sum_{i=1}^{n}(Obs(i)-\overline{Obs})^2}}]^2 \quad (3)$$

여기서 Model(i)는 CMAQ 또는 SVM으로 개선된 모델의 결과를, Obs(i)는 도시대기측정소의 관측값을, 변수 위의 바는 해당 변수의 평균을 의미한다.

진단평가시스템의 기상입력자료는 NCEP의 GFS와 기상청의 UM 자료를 사용하고 있으며 두 경우 모두 예측모델의 결과로 생성되어진다. 이에 따라 CMAQ 모델에서도 예측치를 포함하게 되는데 입력되는 기상자료와 계산에 소요되는 시간을 고려하면 Fig. 3과 같은 시간 주기성을 가지게 된다. 입력되는 기상자료에 따라 CMAQ 모델은 총 95시간 또는 87시간의 계산치를 가지게 되나 계산에 소요되는 시간 때문에 계산대상 첫날 약 6시간 이후에 모델의 계산이 종료되면서 첫날 6시간의 결과는 예측의 기능을 못하게 된다. 하지만 현업에서는 일평균 농도 범위로 대기질 예측이 수행되고 있으므로 본 연구에서도 모델종료 당일부터 총 3일간을 24시간 단위로 구분하여 각각 'day', 'day+1', 'day+2'로 명명하고 예측 일자별로 분석을 수행하였다.

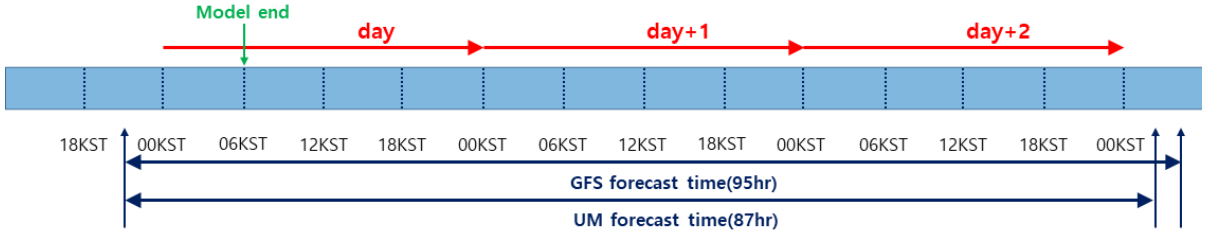

Fig. 3. Schematic diagram of the forecasting time table in the air quality diagnosis and evaluation system.

3. 결과 및 고찰

3.1. 진단평가시스템 적합성 검증

진단평가시스템의 CMAQ 모델결과의 적합성을 확인하기 위하여 2019년에서 2021년간 27개 도시대기측정소 위치에서의 O_3, PM10, PM2.5 예측결과와 실측치 간의 MB, RMSE, R^2를 계산하였다. Table 3은 시간자료를 대상으로 계산한 결과이며 Table 4는 모델치와 실측치의 일평균을 대상으로 계산한 결과를 나타낸다. GFS 기상입력 자료를 사용한 CMAQ 모델의 시간별 예측결과의 설명력은 예측일자별로 O_3 0.30-0.27, PM2.5 0.30-0.24, PM10이 0.20-0.15 순으로 나타났으며 예측기간이 길어질수록 더 낮아지는 경향을 보였다. 이는 모델의 계산시간이 길어질수록 예측된 기상입력자료가 포함하는 불확실성이 더 높아지기 때문으로 판단된다. 모델의 오차 경향을 살펴보면 O_3와 PM2.5는 모델치가 실측치보다 높게 나타나고 있으며 PM10은 과소평가하는 것으로 나타났고 오차율은 O_3 0.017ppm, PM2.5 13ug/m^3 PM10 20ug/m^3으로 계산되었다. R^2와 동일하게 오차율도 예측기간이 길어질수록 높아지는 경향을 보이고 있다(Table 3). 기상청 UM 기상입력 자료를 사용한 CMAQ 모델의 시간별 예측결과의 설명력은 O_3 0.30-0.27, PM2.5 0.28-0.27, PM10이 0.19-0.17 순으로 나타났으며 GFS의 경우와 동일하게 예측기간이 길어질수록 더 낮아지는 경향을 보였다. 모델의 오차 경향도 GFS와 동일한 경향을 보였고 오차율은 O_3 0.019ppm, PM2.5 13ug/m^3 PM10 20ug/m^3으로 계산되었으며 시간별 CMAQ 결과의 정확도는 전반적으로 GFS 기상입력자료가 UM의 경우와 비교하여 다소 높은 것을 알 수 있었다(Table 3). GFS 입력자료를 사용한 일평균의 설명력은 O_3 0.32-0.30, PM2.5 0.48-0.38, PM10 0.40-0.31 순이었으며 예측기간이 길어질수록 기상자료의 정확성이 낮아지면서 감소하는 것으로 나타났다(Table 4). 시간별자료와 동일하게 O_3와 PM2.5는 과대모의, PM10은 과소모의 하는 경향을 보이고 있으며 오차율은 O_3 0.013ppm, PM2.5 9ug/m^3 PM10 15ug/m^3으로 시간자료보다는 감소한 것으로 나타났

다. UM 입력자료를 사용한 경우 일평균의 설명력은 O_3 0.33-0.30, PM2.5 0.44-0.42, PM10 0.38-0.35 순이었으며 GFS의 경우와 동일하게 O_3와 PM2.5는 과대모의, PM10은 과소모의 하는 경향이 나타났다. 오차율은 O_3 0.014ppm, PM2.5 9ug/m^3 PM10 15ug/m^3으로 GFS의 경우와 동일하였으나 시간자료의 경우와 동일하게 GFS가 UM보다 모델결과의 정확도가 증가한 것으로 나타났다. GFS와 UM의 입력자료에 대한 CMAQ 모델의 적합성 분석결과 GFS가 UM보다 개선된 결과를 보이는 것으로 판단되며 이는 입력되는 기상자료 품질의 차이에 의한 것으로 판단된다. 또한 예측결과를 일평균으로 환산한 경우 시간자료가 가지는 변동성이 제거되면서 적합성이 증가하였고 모델의 예측기간이 길어질수록 기상자료의 부정확성이 증가하면서 적합성이 낮아지는 것을 알 수 있었다.

광화학수치모형을 활용하여 시간 또는 일평균 대기질을 예측할 경우 앞에서 살펴본 것처럼 모델의 적합성 문제로 실효성에 의문이 들게 된다. 따라서 대기오염지수나 등급을 활용한 예측방법이 국 내외에서 활용되고 있다[3, 4, 5, 6, 7]. 국내의 경우 통합대기환경지수에 따른 농도 등급을 활용하고 있는데 농도 수준에 따라 '좋음', '보통', '나쁨', '매우나쁨'의 4단계로 구분하고 있다(Table 5, 6). O_3의 경우 일 최고 농도, PM2.5와 PM10은 일평균 농도를 계산하고 해당되는 등급을 활용하여 예보를 수행한다. Table 5, 6은 GFS와 기상청 UM 기상입력 자료를 적용한 CMAQ 예측결과를 농도 범위로 환산하고 실제 관측값의 농도 범위와 일치 여부를 나타낸 것이다. 모델결과의 일치도는 측정소별로 계산하지 않고 부산광역시에서 대기질 예측 대상지역으로 구분한 4개 권역별로 구분하여 계산하였다. Table에서 숫자는 권역별 농도등급의 발생횟수의 합을 의미하며 괄호는 실제로 발생한 농도등급의 횟수에 대한 모델에서 계산한 농도등급의 등급의 적중률을 의미한다. O_3의 경우 '보통' 등급에서 GFS 98%, UM 94%로 가장 높은 적중률을 보였고 PM2.5와 PM10은

Table 3. Performance of hourly predicted O3, PM10, and PM2.5 from CMAQ model using GFS and UM meteorological input data

		MB			RMSE			R2		
		day	day+1	day+2	day	day+1	day+2	day	day+1	day+2
GFS	O_3	0.006	0.007	0.007	0.017	0.018	0.018	0.30	0.28	0.27
	PM10	-9	-10	-10	20	21	21	0.20	0.18	0.15
	PM2.5	1.1	0.6	0.2	13	13	13	0.30	0.28	0.24
UM	O_3	0.007	0.007	0.007	0.019	0.019	0.019	0.30	0.29	0.27
	PM10	-9	-9	-9	20	20	20	0.19	0.17	0.18
	PM2.5	1.0	1.0	0.7	13	13	13	0.28	0.27	0.27

Table 4. Performance of daily mean predicted O3, PM10, and PM2.5 from CMAQ model using GFS and UM meteorological input data

		MB			RMSE			R^2		
		day	day+1	day+2	day	day+1	day+2	day	day+1	day+2
GFS	O_3	0.006	0.007	0.007	0.013	0.013	0.013	0.32	0.31	0.30
	PM10	-9	-10	-10	15	16	16	0.40	0.37	0.31
	PM2.5	1.0	0.6	0.2	9	9	10	0.48	0.45	0.38
UM	O_3	0.008	0.008	0.008	0.014	0.015	0.015	0.33	0.31	0.30
	PM10	-9	-9	-9	15	16	15	0.38	0.35	0.37
	PM2.5	1.0	1.0	0.6	9	9	9	0.44	0.42	0.44

'좋음' 등급에서 가장 높은 적중률을 보였다. 연구대상 기간 중 황사 발생일과 전후 일을 분석에서 제외하여 상대적으로 고농도 발생 일이 감소한 결과로 판단된다. 상대적으로 고농도에 해당되는 '나쁨' 이상의 적중률은 GFS의 경우 39(O_3)-56%(PM2.5) 였으며 UM의 경우 48(O_3)-49%(PM2.5) 로 나타나 현업에서 예측을 위한 기본 자료로 사용하기에는 무리가 없을 것으로 판단된다. 대기오염 진단평가시스템에서 사용하고 있는 두 종류의 기상자료에 대한 O_3, PM2.5, PM10의 예측결과를 관측자료와 비교한 결과 고농도 등급의 적중률은 현업에서 활용가능 한 것으로 판단되나, 1km 상세격자별로 생성되는 시간별 예측결과를 그대로 사용

Table 5. Performance of prediction by concentration grade using GFS meteorological input data

			GFS				
			very unhealthy	unhealthy	moderate	good	total
OBS	≧0.151	very unhealthy	0(0%)	5	1	0	6
	≧0.091	unhealthy	3	104(39%)	162	0	269
	≧0.031	moderate	0	74	3,844(98%)	2	3,920
	≧0	good	0	0	75	2(3%)	77
	O_3(ppm)	total	3	183	4,082	4	4,272
	≧76	very unhealthy	0(0%)	1	0	0	1
	≧36	unhealthy	0	133(56%)	92	11	236
	≧16	moderate	0	177	1159(69%)	350	1686
	≧0	good	0	9	410	1614(79%)	2033
	PM2.5(ug/m^3)	total	0	320	1661	1975	39856
	≧151	very unhealthy	0(0%)	0	0	0	0
	≧81	unhealthy	0	0(0%)	27	5	32
	≧31	moderate	0	0	577(37%)	988	1565
	≧0	good	0	0	81	2278(97%)	2359
	PM10(ug/m^3)	total	0	0	685	3271	3956

Table 6. Performance of prediction by concentration grade using UM meteorological input data

			UM				
			very unhealthy	unhealthy	moderate	good	total
OBS	≧0.151	very unhealthy	0(0%)	5	1	0	6
	≧0.091	unhealthy	1	131(48%)	142	0	274
	≧0.031	moderate	0	219	3728(94%)	4	3951
	≧0	good	0	0	68	9(12%)	77
	O_3(ppm)	total	1	355	3939	13	4308
	≧76	very unhealthy	0(0%)	1	0	0	1
	≧36	unhealthy	0	113(49%)	105	14	232
	≧16	moderate	0	183	1178(69%)	342	1703
	≧0	good	0	18	492	1542(75%)	2052
	PM2.5(ug/m^3)	total	0	315	1775	1898	3988
	≧151	very unhealthy	0(0%)	0	0	0	0
	≧81	unhealthy	0	0(0%)	28	4	32
	≧31	moderate	0	0	576(37%)	998	1574
	≧0	good	0	0	85	2297(96%)	2382
	PM10(ug/m^3)	total	0	0	689	3299	3988

하기에는 적합성 변수가 낮아 무리가 있을 것으로 판단이되며 따라서 기계학습을 활용한 모델결과 개선의 필요성을 확인 할 수 있었다.

3.2. CMAQ 결과 개선을 위한 SVM의 적용

O_3, PM10, PM2.5 시간자료에 대한 낮은 적합성을 개선하기 위하며 본 연구에서는 CMAQ 모델링 결과에 SVM을 적용하였다. Table 5는 CMAQ 시간별 결과에 SVM을 적용하고 모델의 적합성 변수를 계산한 결과이다. 모델의 설명력은 O_3 0.69, PM2.5 0.65, PM10 0.55로 시간별 CMAQ 모델의 결과와 비교하면 상당히 개선되는 것으로 나타났다. 오차경향을 살펴보면 O_3, PM10, PM2.5 모두 약하게 과소모의하고 있으며 오차율도 O_3 0.010ppm, PM10 12.9ug/m^3, PM2.5 7.5ug/m^3으로 단일모델의 0.017, 22, 14와 비교하여 상당히 개선되고 있다.

SVM을 CMAQ 모델의 결과에 적용하면 CMAQ의 결과를 직접 사용하는 경우보다 모델의 적합성 변수들

Table 7. Performance of hourly predicted O3, PM10, and PM2.5 from CMAQ+SVM model

	MB			RMSE			R2		
	day	day+1	day+2	day	day+1	day+2	day	day+1	day+2
O_3	-0.000	-0.000	-0.000	0.010	0.010	0.010	0.69	0.68	0.67
PM10	-1.4	-1.4	-1.5	12.8	12.9	13.3	0.55	0.54	0.51
PM2.5	-0.9	-0.9	-0.9	7.5	7.6	7.7	0.65	0.65	0.64

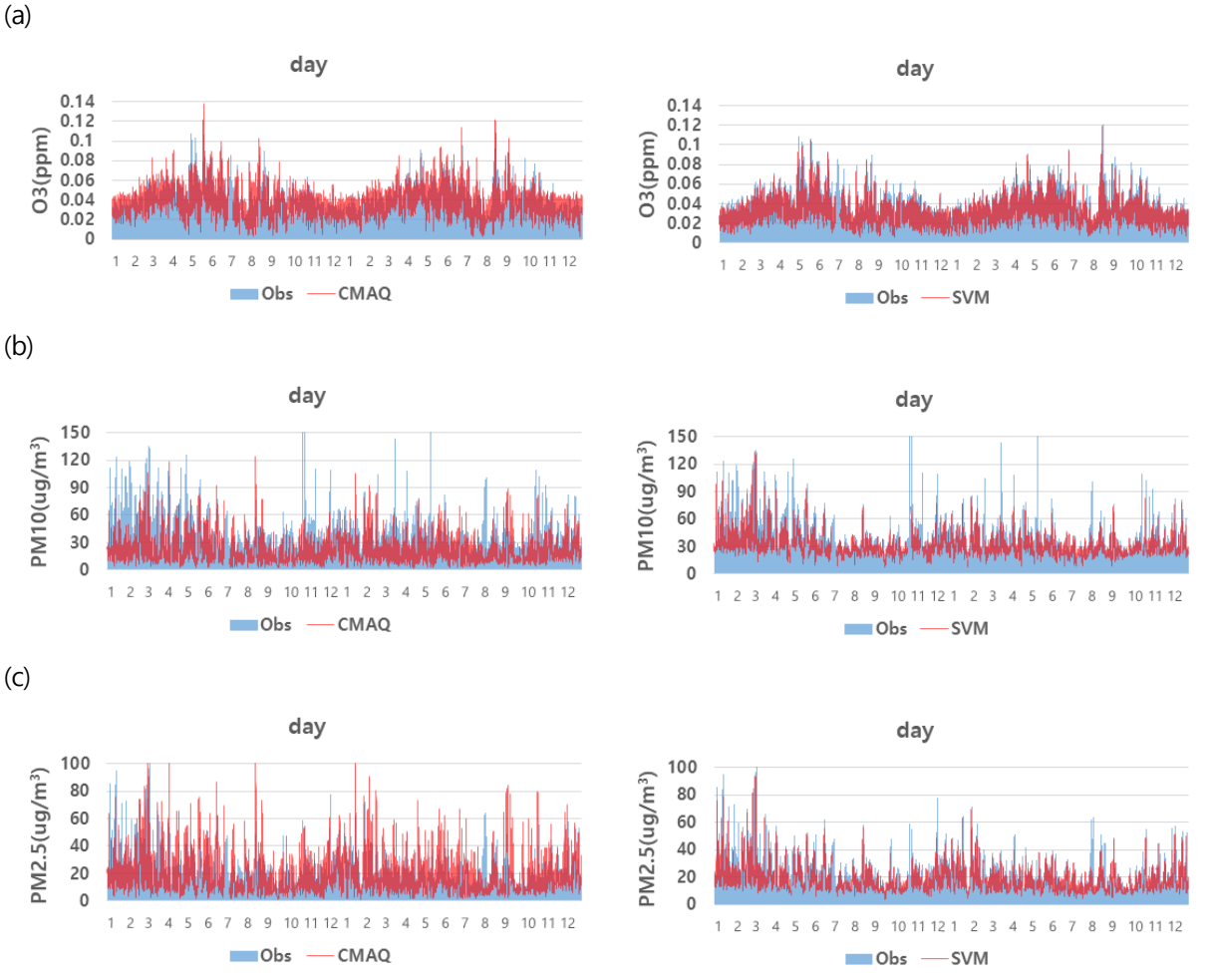

Fig. 4. Comparison of hourly predicted O3, PM10 and PM2.5 time series between CMAQ and SVM.

이 개선되는 것을 확인하였다. 실제로 시계열자료의 변동성이 잘 반영되는지 확인하기 위하여 'day' 케이스의 O_3, PM10, PM2.5의 시계열 자료를 Fig. 4에 제시하였다. 그림에서 왼쪽은 CMAQ 모델, 오른쪽은 SVM을 적용한 경우이다. SVM을 적용하면 CMAQ 결과보다 과대 또는 과소모의되는 범위가 줄어드는 것이 확인되며 이로 인하여 RMSE값이 개선된 것으로 판단된다. 예측값의 편향성을 보여주는 MB도 크게 개선되는 것으로 나타났는데 그림에서도 관측값의 패턴과 거의 일치하는 것을 확인할 수 있다. 모델의 적합성 변수와 시계열 패턴의 변화를 통하여 SVM을 적용할 경우 관측값과의 편향성과 오차율이 크게 개선이되며 시계열의 매시간 변동성을 잘 따라가도록 CMAQ의 결과가 개선되는 것을 확인할 수 있었다.

3.3. 지점별 CMAQ 개선 결과

SVM 적용에 따른 CMAQ 모델결과의 개선 효과를 상세하게 보기 위하여 지점별로 모델 적합성 변수의 변화를 Fig. 5-7에 제시하였다. Fig. 5는 지점별 O_3의 모델 적합성 변수의 변화를 나타낸 그림이다. 모델의 설명력을 나타내는 R^2는 명장동측정소 0.50에서 0.73으로 개선되었으며 명지동측정소에서 가장 높은 0.25에서 0.93으로 개선되었다. 명지동측정소는 2020년 9월 신설된 측정소로 모델결과와 비교할 수 있는 관측값의 수가 적어서 상대적으로 높은 개선효과가 나타난 것으로 판단된다. 2019년부터 운영하고 있는 측정소 중에서 가장 개선효과가 높은 지점은 대신동측정소로 0.19에서 0.51로 개선되었다. 모델의 오차율을 나타내는 RMSE도 전지점에서 개선되는 것으로 나타났는데 명장동측정소에서 0.005ppm(0.014→0.009ppm) 개

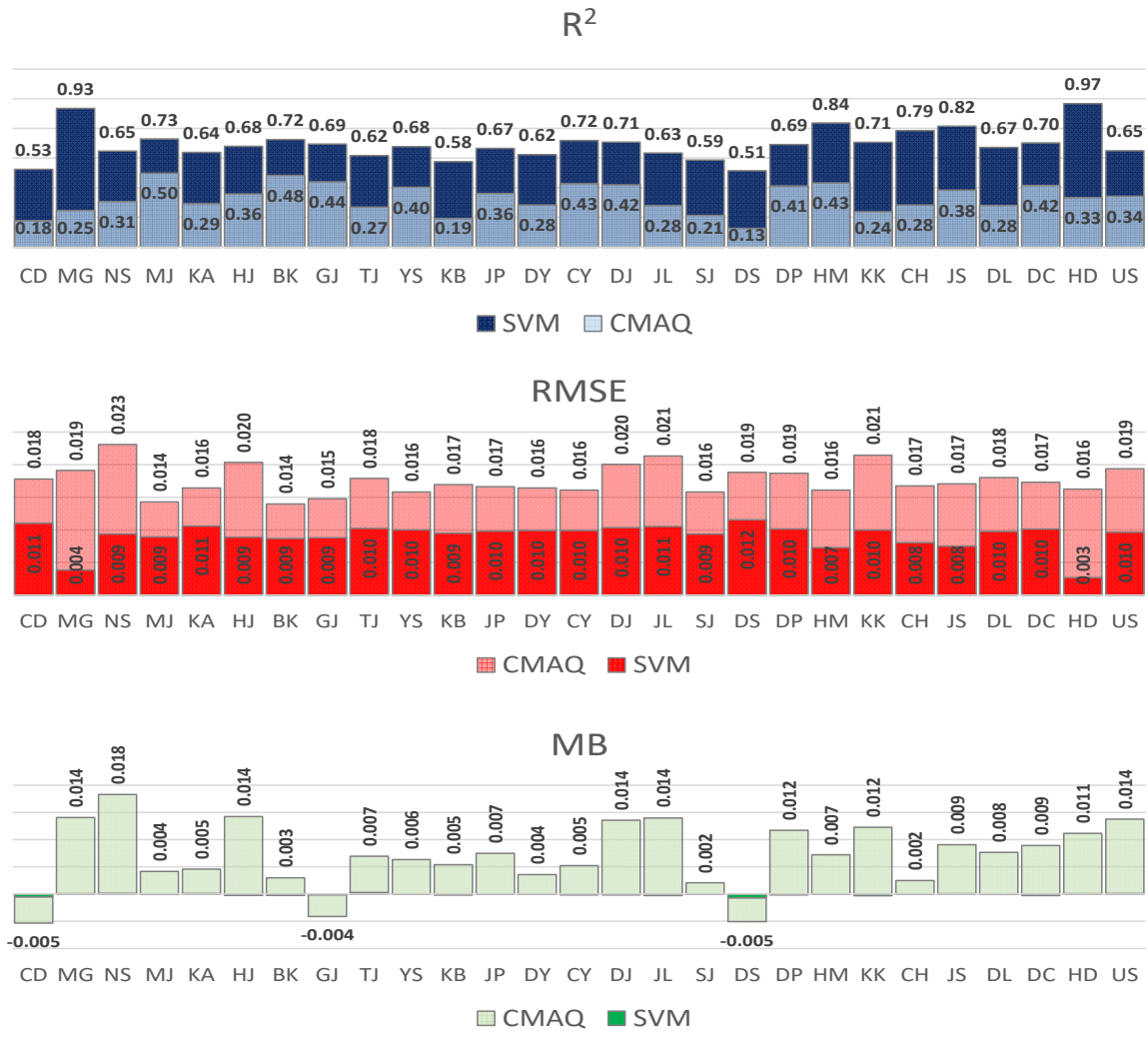

Fig. 5. Differences of O_3 performance between CMAQ and SVM by air quality stations.

선되었고 명지동측정소에서 0.015ppm(0.019→0.004ppm)으로 가장 개선효과가 크게 나타났다. 2019년 이전부터 운영된 측정소 중에는 녹산동측정소에서 0.014ppm(0.023→0.009ppm)으로 가장 높았다. CMAQ 모델의 결과는 대부분의 지점에서 실측치를 과대 모의 하였으나 SVM을 적용하면 전 지점에서 편향성이 감소되는 것을 확인할 수 있다. 지점간 개선효과의 차이는 CMAQ 모델의 입력자료와 실측치의 변화에 영향을 주는 주변 요인들의 차이가 반영된 것으로 판단이 되며 SVM을 적용할 경우 모델의 반복학습을 통하여 이러한 차이를 최소화 하는 효과가 있는 것으로 판단된다.

Fig. 6은 지점별 PM10의 모델 적합성 변수의 변화를 나타낸 그림이다. PM10의 지점별 설명력은 명장동측정소에서 0.031(0.21→0.52), 명지동측정소에서 0.83(0.07→0.90)으로 가장 큰 차이를 보이는 것으로 나타났다. 2019년 이전부터 운영된 측정소 중에서는 학장동, 태종대, 대신동에서 0.37로 개선효과가 가장 크게 나타났다. O_3와 동일하게 RMSE도 전지점에서 개선되는 것으로 나타났는데 당리동측정소에서 6.9ug/m^3(17.5→10.6ug/m^3), 명지동측정소에서 최대 18.1ug/m^3(24.0→5.8ug/m^3)로 개선되었다. 당리동측정소는 2019년 5월 신설되어 명지동과 동일하게 상대적으로 모델치와 비교할 수 있는 관측치수가 부족

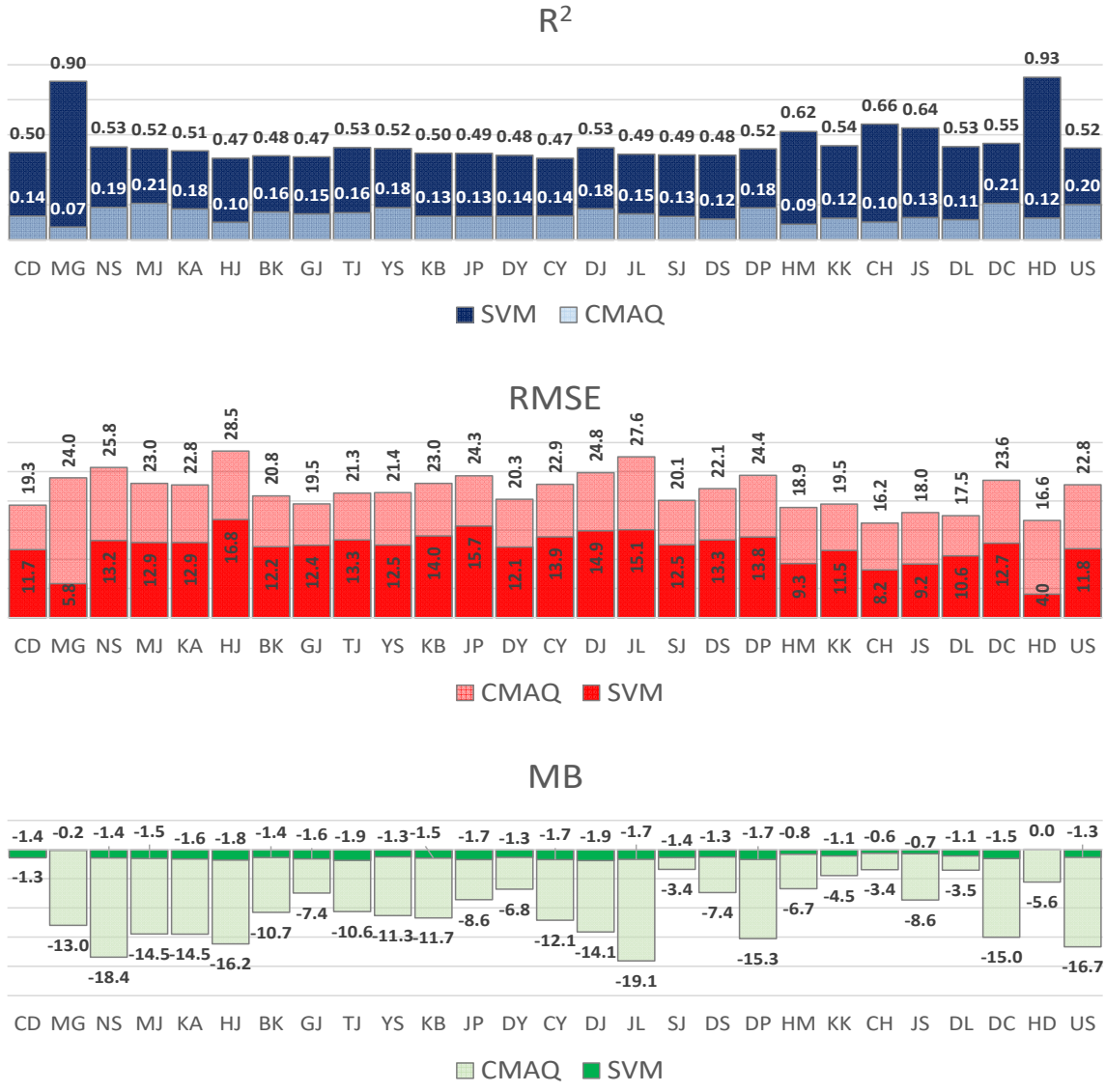

Fig. 6. Differences of PM10 performance between CMAQ and SVM by air quality stations.

하며 2019년 이전부터 운영된 측정소 중에는 기장읍에서 7.1ug/m^3(19.5→12.4ug/m^3), 장림동과 녹산동 측정소에서 12.5ug/m^{3}의 개선효과가 나타났다. CMAQ 모델의 PM10 예측결과는 전지점에서 관측치를 과소모의 하였으며 SVM을 적용한 결과도 모델치가 과소모의 하였으나 그 차이는 상당히 감소한 것을 알 수 있었다.

Fig. 7은 CMAQ와 SVM간의 지점별 PM2.5의 모델 적합성 변수의 변화를 나타낸 그림이다. PM2.5의 지점별 설명력 변화의 최소값은 기장읍, 연산동, 덕포동 측정소에서 0.30이고 최대값은 명지동, 회동동측정소에서 0.67로 나타났다. 회동동측정소는 2020년 10월 신설된 측정소이며 2019년 이전부터 운영된 측정소 중에는 광복동에서 0.36(0.23→0.59) 가장 높게 나타났다. 모델의 오차율을 나타내는 RMSE도 전지점에서 개선되는 것으로 나타났는데 용수리측정소에서 4.9ug/m^3(11.9→7.0ug/m^3), 회동동측정소에서 9.6ug/m^3(11.4→1.9ug/m^3)까지 지점별 개선효과가 나타났다. 2019년 이전부터 운영된 측정소 중에는 수정동측정소에서 최대 7.7ug/m^3(14.5→6.8ug/m^3)로 개선되었다. CMAQ 모델의 PM2.5 예측결과는 지점별로 편차가 차이가 있었으나 SVM을 적용하면 전지점에서 관측치를 과소모의 하게된다. 하지만 그 정도는 줄어들면서 모델의 정합도가 증가하는 것을 알 수 있다.

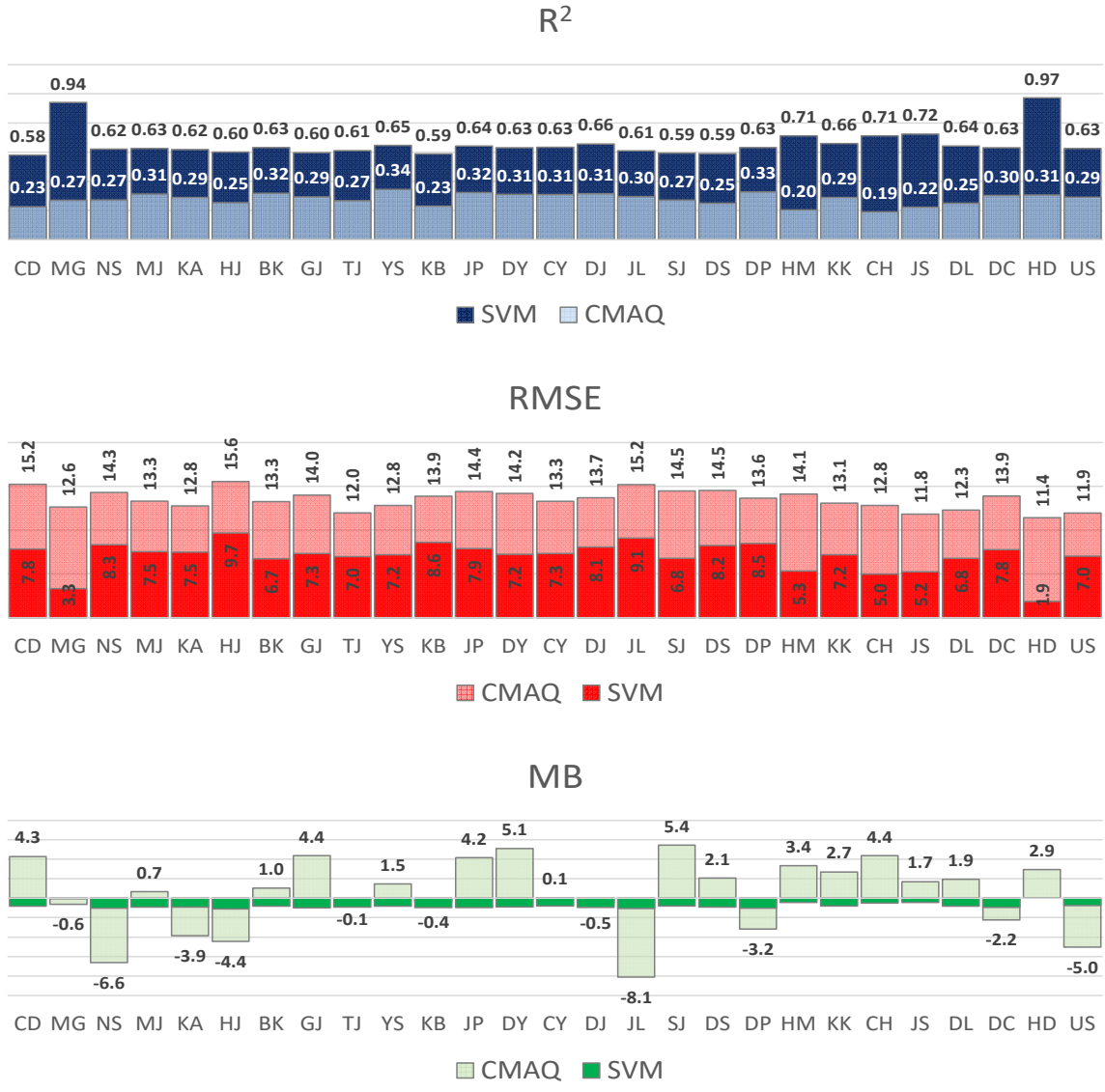

Fig. 7. Differences of PM2.5 performance between CMAQ and SVM by air quality stations.

4. 결론 및 제언

본 연구는 부산광역시 보건환경연구원에서 실시간으로 운영중인 CMAQ 모델링 시스템의 O_3, PM10, PM2.5의 시간별 예측결과를 개선하는 것을 목적으로 한다. 진단평가 시스템은 광화학 수치모델 CMAQ을 기본으로 하며 모델의 계산종료 시점에 3일의 예측치를 포함하고 있다. 모델의 입력자료와 물리, 화확적인 계산과정에서 기본적으로 불확실성이 포함되어 있으며 이를 개선하기 위하여 머신러닝 기법인 SVM을 시간별 O_3, PM10, PM2.5 모델결과에 적용하였다.

1. 진단평가시스템 CMAQ 모델의 시간별 예측결과의 설명력은 O_3 0.30, PM2.5 0.30, PM10이 0.20 순으로 나타났으며 예측기간이 길어질수록 더 낮아지는 경향을 보였다. 모델의 오차 경향을 살펴보면 O_3와 PM2.5는 모델치가 실측치보다 높게 나타나고 있으며 PM10은 과소평가하는 것으로 나타났고 오차율은 O_3 0.017ppm, PM2.5 13ug/m^3, PM100 20ug/m^3으로 계산되었으며 예측기간이 길어질수록 높아지는 경향을 보이고 있다.
2. 시간별 CMAQ 결과에 SVM을 적용하면 모델의 설명력은 O_3 0.69, PM2.5 0.65, PM10 0.55로 시간별 CMAQ 모델의 결과와 비교하면 상당히 개선되었으며 오차경향을 살펴보면 O_3, PM10, PM2.5 모두 실측치를 약하게 과소모의하고 있으며 오차율도 O_3 0.010ppm, PM10 12.9ug/m^3, PM2.5 7.5ug/m^3으로 단일모델의 0.017, 22, 14와 비교하여 상당히 개선되었다.
3. SVM 적용에 따른 CMAQ 모델결과의 개선 효과를 지점별로 살펴보면 O_3의 설명력(R^2)은 0.23에서 0.69까지 개선되었고 오차율(RMSE) 0.015에서 0.005ppm으로 개선되었다. PM10의 경우 설명력(R^2)은 0.31에서 0.83까지 개선되었고 오차율(RMSE)은 18.13에서 6.89ug/m^3까지 개선되었다. 지점별 PM2.5의 설명력(R^2)은 0.30에서 0.67까지, 오차율(RMSE)은 9.55에서 4.92 ug/m^3 까지 개선되어 전지점에서 모델의 적합성을 향상시키는 것을 확인하였다.

5. 참고문헌

1. National Oceanic and Atmospheric Administration, National Weather Service, https://airquality.weather.gov/sectors/conus.php(2021).
2. Government of Canada, Regional Air Quality Deterministic Prediction System, https://weather.gc.ca/aqfm/index_e.html(2021).
3. Department for Environment Food & Rural Affairs, UK AIR, https://uk-air.defra.gov.uk/forecasting/?day=2#forecast_map(2021).
4. Ministry of Earth Science, Govt. of India, Indian Institute of Tropical Meteorology, System of Air Quality and Weather Forecasting And Research, http://safar.tropmet.res.in/map_data.php?for=current&city_id=1(2021).
5. 국립환경과학원 대기질통합예보센터, 대기정보예보, https://www.airkorea.or.kr/web/dustForecast?pMENU_NO=113(2021).
6. 부산일보, http://www.busan.com/view/busan/view.php?code=20180130000328(2018).
7. 부산광역시 보건환경정보공개시스템, 대기질 진단평가, https://heis.busan.go.kr/environmental/air006.aspx(2021).
8. Kim, J. and Jang, Y. K., “Uncertainty Assessment for CAPSS Emission Inventory by DARS”, *Journal of Korean Society for Atmospheric Environment*, 30(1), pp.26-36(2014).
9. Jo, Y. J., Lee, H. J., Chang, L. S. and Kim, C. H., “Sensitivity Study of the Initial Meteorological Fields on the PM10 Concentration Predictions Using CMAQ Modeling”, *Journal of Korean Society for Atmospheric Environment*, 33(6), pp.554-569(2017).
10. Kitayamaa, K., Morinoa, Y., Yamajib, K. and Chatania, S., “Uncertainties in O_3 concentrations simulated by CMAQ over

Japan using four chemical mechanisms", *Atmospheric Environment*, 198, pp.448-462(2019).
11. Choi, D. R. and Koo, Y. S., "An Evaluation of the Influence of Boundary Conditions from GEOS-Chem on CMAQ Simulations over East Asia", *Journal of Korean Society for Atmospheric Environment*, 29(2), pp.186-198(2013).
12. Rybarczyk, Y. and Zalakeviciute, R., "Machine learning approaches for outdoor air quality modelling: a systematic review", *Appl Sci*, 8(12), p2570(2018).
13. Joharestani, M. Z., Cao, C., Ni, X., Bashir, B. and Talebiesfandarani, S., "PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data", *Atmosphere*, 10(7), p373(2019).
14. Dutta, A. and Jinsart, W., "Air Pollution in Indian Cities and Comparison of MLR, ANN and CART Models for Predicting PM10 Concentrations in Guwahati, India", *Asian Journal of Atmospheric Environment*, 15(1), pp.1-26(2021).
15. Shahriar, S. A., Kayes, I., Hasan, K., Salam, M. A. and Chowdhury, S., "Applicability of machine learning in modeling of atmospheric particle pollution in Bangladesh", *Air Quality, Atmosphere & Health*, 13, pp.1247-1256(2020).
16. Madhavi, A. E., Naresh, S., Kim, N. D. and Jennifer, A. S., "Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis", *Atmospheric Pollution Research*, 5, pp.696-708(2014).
17. Goulier, L., Paas, B., Ehrnsperger, L. and Klemm, O., "Modelling of Urban Air Pollutant Concentrations with Artificial Neural Networks Using Novel Input Variables", *International Journal of Environmental Research and Public Health*, 17(6), p2025(2020).
18. Lim, J. M., "An Estimation Model of Fine Dust Concentration Using Meteorological Environment Data and Machine Learning", *Journal of Information Technology Services*, 18(1), pp.173-186(2019).
19. Cha, J. W. and Kim, J. Y., "Development of Data Mining Algorithm for Implementation of Fine Dust Numerical Prediction Model", *Journal of the Korea Institute of Information and Communication Engineering*, 22(4), pp.595-601(2018).
20. Cho, K. H., Lee, B. Y., Kwon, M. H. and Kim, S. C., "Air Quality Prediction Using a Deep Neural Network Model", *Journal of Korean Society for Atmospheric Environment*, 35(2), pp.214-225(2019).
21. Sayeed, A., Choi, Y., Eslami1, E., Jung, J., Lops, Y., Salman, A. K., Lee, J. B., Park, H. J. and Choi, M. H., "A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance", *Scientific Reports*, p11, 10891(2021).
22. NCAR Research Data Archive, NCEP GFS 0.25 Degree Global Forecast Auxiliary Grids Historical Archive, https://rda.ucar.edu/datasets/ds084.3/#description(2021).
23. 국가기상슈퍼컴퓨터센터, https://www.kma.go.kr/aboutkma/intro/supercom/index.jsp(2021).
24. WRF MODEL USERS'PAGE, https://www2.mmm.ucar.edu/wrf/users/(2021).
25. Community Modeling and Analysis System, Sparse Matrix Operator Kerner Emissions Modeling System, https://www.cmascenter.org/smoke/(2021).
26. MEICModel, Tracking Anthropogenic Emissions in China, http://meicmodel.org/(2021).
27. Regional Emission inventory in ASia Data Download Site, https://www.nies.go.jp/REAS/index.html(2021).
28. Ohara, T., Akimoto, H., Kurokawa, J., Horii,

N., Yamaji, K., Yan, X. and Hayasaka, T., “An Asian emission inventory of anthropogenic emission sources for the period 1980-2020”, *Atmos. Chem. Phys.*, 7, pp.4419-4444(2007).
29. EPA, Biogenic Emission Inventory System(BEIS), https://www.epa.gov/air-emissions-modeling/biogenic-emission-inventory-system-beis(2021).
30. Washington State University, Model of Emissions of Gases and Aerosols from Nature (MEGAN), http://bioearth.wsu.edu/megan_model.html(2021).
31. EPA, CMAQ: The Community Multiscale Air Quality Modeling System, https://www.epa.gov/cmaq(2021).
32. Byun, D. W. and Ching, J. K. S., “Science algorithms of the EPA Models-3 Community Multiscale Air Quality(CMAQ) Modeling System”, *U.S. Environmental Protection Agency(US EPA)*, EPA/600/R-99/030(1999).
33. 부산광역시 보건환경정보공개시스템, 실시간측정자료, https://heis.busan.go.kr/environmental/air001.aspx(2021).
34. Cortes, C. and Vapnik, V., “Support-Vector Networks”, *Machine Learning*, 20, pp.273-297(1995).
35. Son, S. H. and Kim, J. S., “Evaluation and Predicting PM10 Concentration Using Multiple Linear Regression and Machine Learning”, *Korean Journal of Remote Sensing*, 36(6-3), pp.1711-1720(2020).
36. Pourghasemi, H. R., Jirandeh, A. G., Pradhan, B., Xu, C. and Gokceoglu, C., “Landslide susceptibility mapping using support vector machine and GIS at the Golestan Province, Iran”, *Journal of Earth System Science*, 122(2), pp.349-369(2018).
37. Chen, H. L., Yang, B., Liu, J. and Liu, D. Y., “A support vector machine classifier with rough setbased feature selection for breast cancer diagnosis”, *Expert Systems with Applications*, 38(7), pp.9014-9022(2011).
38. Yanga, X., Wua, Q., Zhaob, R., Cheng, H., He, H., Ma, Q., Wang, L. and Luo, H., “New method for evaluating winter air quality: PM2.5 assessment using Community Multi-Scale Air Quality Modeling (CMAQ) in Xi'an”, *Atmospheric Environment*, 211, pp.18-28(2019).
39. National Air Emission Inventory and Research Center, https://www.air.go.kr/en-main.do(2021).