2021년 한국환경영향평가학회 추계학술대회

환경 빅데이터 플랫폼
실내공기질 빅데이터 센터

빅데이터와 딥러닝을 이용한 저가형 PM2.5 센서의 구간별 보정 알고리즘 개발

KAIST 건설 및 환경공학과
류제완

2021.10.28

KAIST

Introduction – 센터 소개

1 자체 생산 빅데이터

기존 운영 학내 테스트베드
KAIST 강의실, 지원시설, 학내 유치원, 학생식당

빅데이터 수집 테스트베드 선정
아파트, 카페, 학교, 병원, 키즈카페 등등

실내 공기질, 인간 행동 IoT 센서 노드
미세먼지 / 행동감지 / CO_2 / 온습도 / 압력 / 풍속

기타 센서 노드
스마트에너지미터 / 스마트 콘센트

연구 차별화 전략
- 세계최고수준 전처리 알고리즘
- 초경량화를 통한 엣지 탑재
- 단순 DB 구축이 아닌 센터로 확장 구축

LoRa/Wifi 통신망 구축

실내 공기질 엣지 게이트웨이
컴퓨팅 모듈 / 저장디바이스 / CPU

로컬에서 1차 판단

IoT 누락 복원 알고리즘 탑재

초경량 이상치 탐지 알고리즘

표준화, 규격화된 IoT 샘플 빅데이터 전송

엣지용 AI 알고리즘 배포 실증테스트베드 시범 적용

전체 개요

환경 빅데이터 공유
플랫폼 서비스 연계

데이터 전처리 기술 공유
데이터 표준 체계 공유

2 환경 빅데이터 플랫폼 연계

환경 플랫폼 및 센터와의 긴밀한 연계

- 물/환경 (K-water)
- 자연/생태 (국립생태원)
- 지질/재해 (한국지질자원연구원)
- 환경영향/저감 (한국환경정책평가연구원)
- 스마트 그리드 (아이렉스넷)
- 화학물질/제품 (그린에코스)
- 실외 미세먼지 (지디에스컨설팅)
- 기상/기후 (한국기상산업기술원)

OCF 기반 빅데이터 공유
행동인자, 피드백 공유

빅데이터 분석 결과 공유
실내 공기질 빅데이터 유통 및 거래

3 실내 공기질 빅데이터 센터

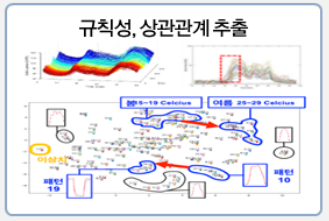
빅데이터 수집 레이어

- 온톨로지 학습기반 데이터 자동 라벨기
- OCF JSON 스키마 국제 IoT 표준화 모델
- 이상신호 데이터 클라우드 기반 정제

빅데이터 저장 레이어 (AWS/Hadoop)

공공 데이터	행정 데이터	IoT 수집 데이터
실외 기상기후	공기청정기 설치 현황 데이터	플랫폼 내 타 센터 DB
건물 단위 전력 사용량	학교 시설 물 신축/공사	실내 공기질 DB
고농도 미세먼지 법령	미세먼지 관련 민원 데이터	열환경, 소음 및 안전 DB

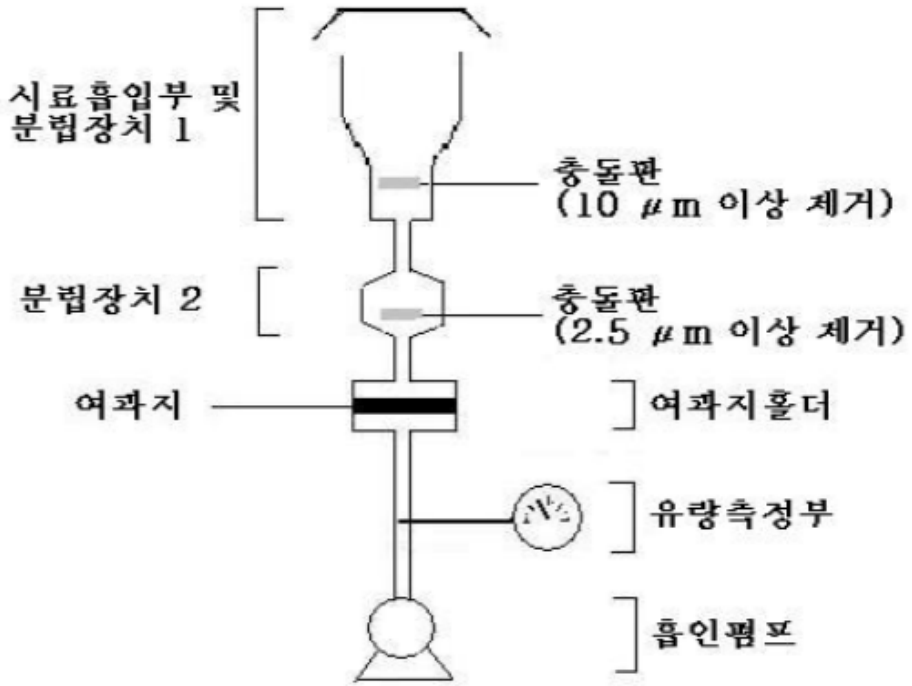
빅데이터 프로파일링



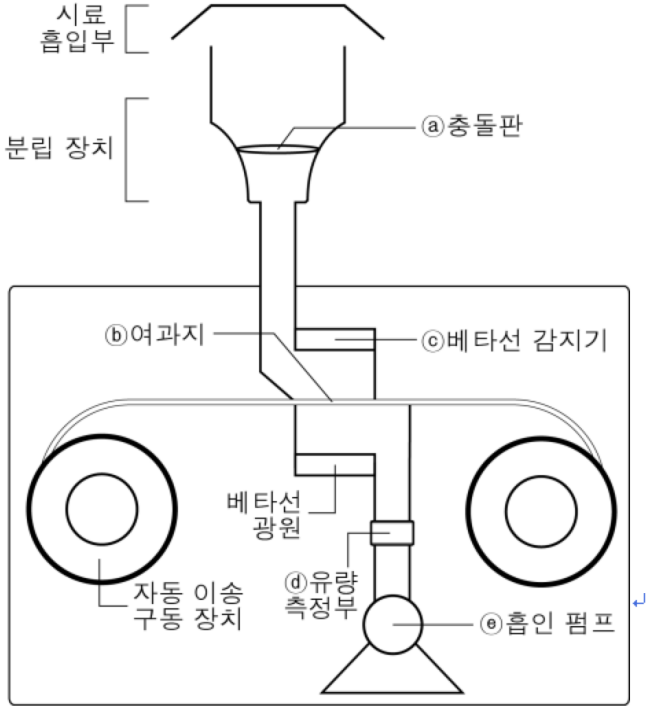
빅데이터 분석 AI 모델 핵심기술

- 동적 모델 RDNN 학습 기반 실내 공기질 최적 제어 의사결정론 개발
- KNN, LSTM 기반 장단기 공조·청정 설비 플래닝
- 엣지 컴퓨팅 환경을 위한 데이터 처리학습 모델 경량화, 최적화

Introduction – 연구배경



Gravimetric method



Beta-ray absorption method

- 넓은 공간
- 고비용
- 시료 건조 과정

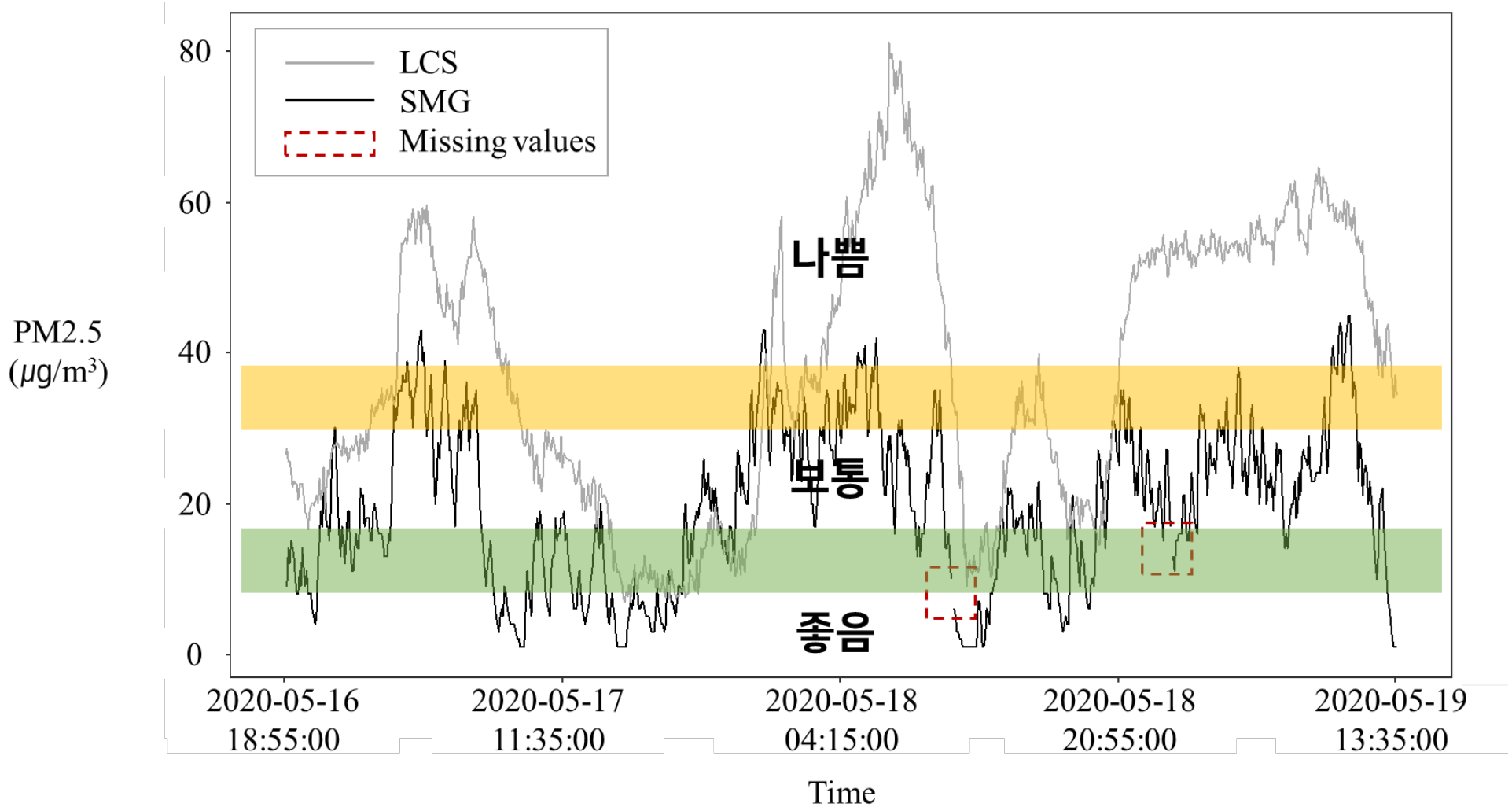
Introduction – 연구배경

- 적외선 산란 방식의 저가형 미세먼지 센서
  - 시료를 건조하는 과정이 없어 습도에 민감
  - 먼지, 돌풍, 담배연기 등 국지적인 이벤트에 민감
  - 오랜 시간 사용하면서 흡기 팬에 먼지가 쌓임
  - 제조 과정에서 보정 작업이 이루어지지만 보정 환경과 실제 사용하는 환경의 차이로 인해 미세먼지의 농도를 overestimate하는 경우가 많음
  - → 사용하는 현장에서 재보정이 필요

Methodology – 데이터 수집 및 전처리

- 대전 보건환경연구원
- Plantower 사의 PMS 7003M 센서
- Thermo Scientific 사의 5014i Beta Continuous Ambient Particulate Monitor 제품 사용
- 센서는 샘플러와 동일한 위치에 설치
- 비와 직사광선을 비하면서 바닥으로부터 충분히 띄워 센서 근처의 공기 흐름을 원활하게 유지
- 데이터 수집 기간: 2020년 5월~2020년 8월 10초 간격으로 데이터 수집
- 5분 간격으로 평균을 내어 총 29,195개의 관측치

Methodology – 데이터 수집 및 전처리

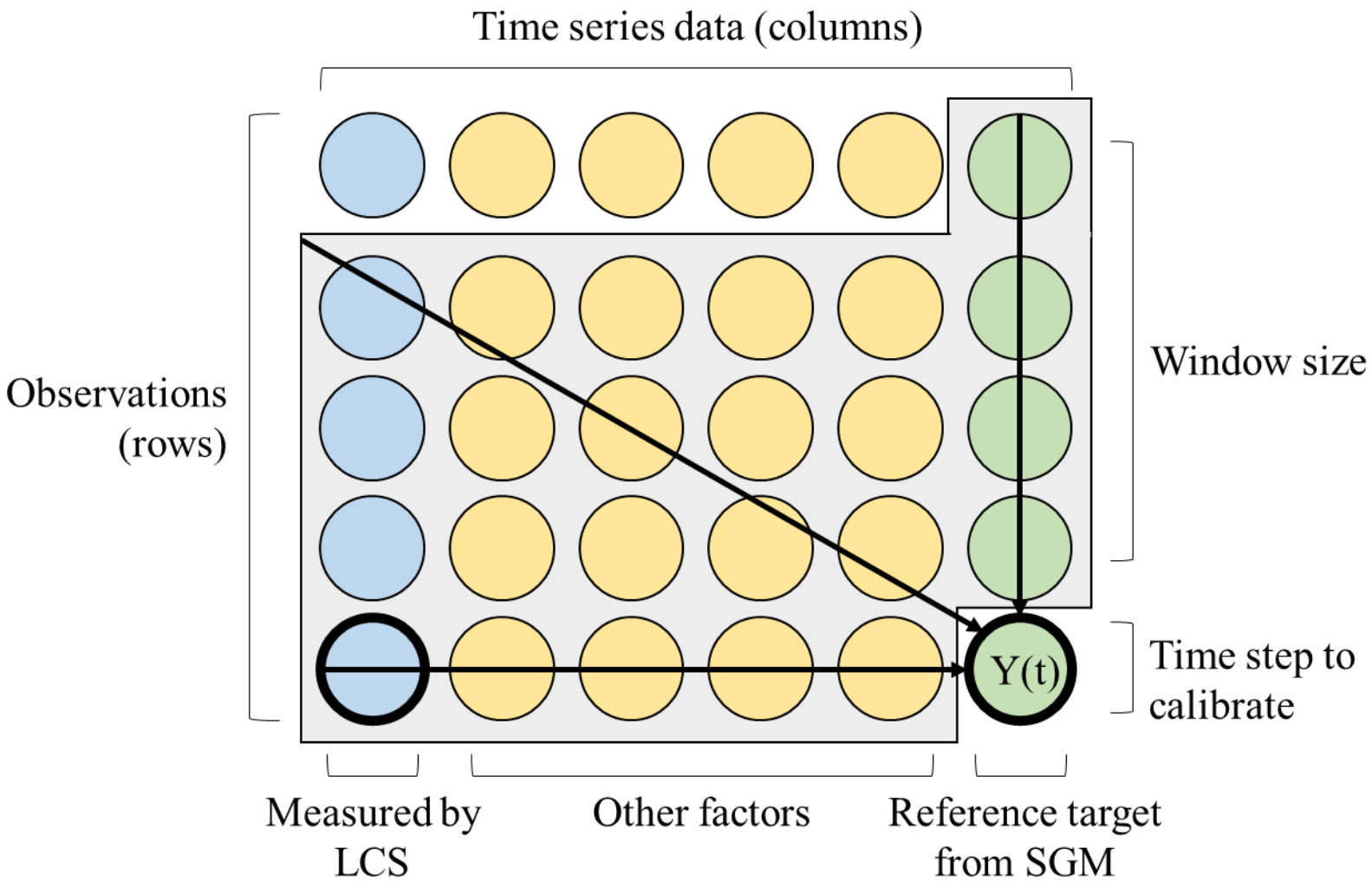


Methodology – 데이터 수집 및 전처리

Data	Units	Mean	Median	Standard deviation	minimum	maximum
PM2.5 (LCS)	㎍/m3	24.63	22.09	19.03	0.00	121.62
PM2.5 (SMG)	㎍/m3	12.71	10.00	10.99	0.00	118.00
CO2	ppm	433.95	425.68	25.27	400.26	550.00
VOC	ppm	673.40	519.29	56.60	9.89	6172.90
SO2	ppm	0.00	0.00	0.01	0.00	0.47
CO	ppm	0.38	0.3	0.57	0.10	40.50
O3	ppm	0.04	0.03	0.02	0.00	0.42
NO2	ppm	0.01	0.01	0.01	0.00	0.05
Temperature	℃	24.63	24.68	3.86	12.70	35.32
Relative humidity	%	82.09	87.08	15.40	26.50	96.90
Wind speed	m/s	1.40	1.14	1.08	0.00	8.48

$$X(t) = \frac{max\left(X(t)\right) - X(t)}{max\left(X(t)\right) - min\left(X(t)\right)}, X(t)\ is\ \text{input for LSTM model}$$

Methodology – 학습 데이터 구조



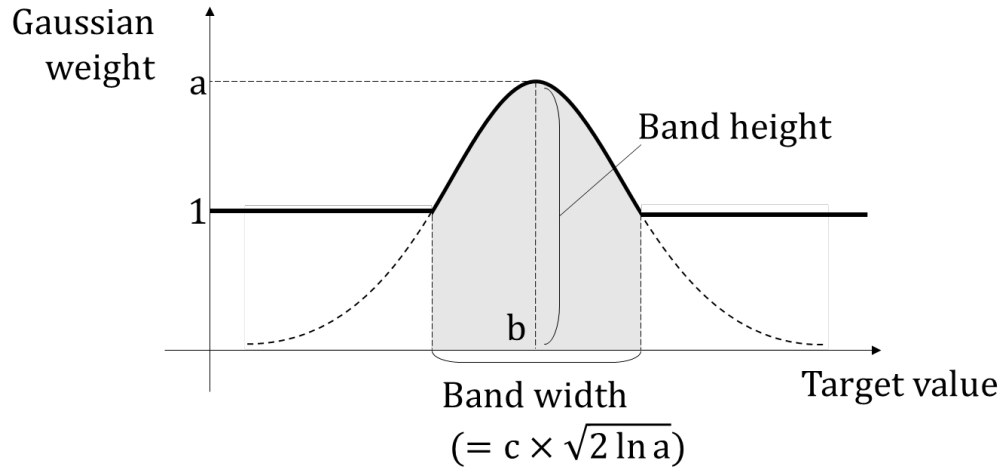
Methodology – Gaussian weighted loss function

- 타겟값의 구간에 따른 dynamic weight를 도입, 선택적으로 보정 능력을 향상 (과속카메라 등)
- Symmetric, decreasing to tail → Gaussian function

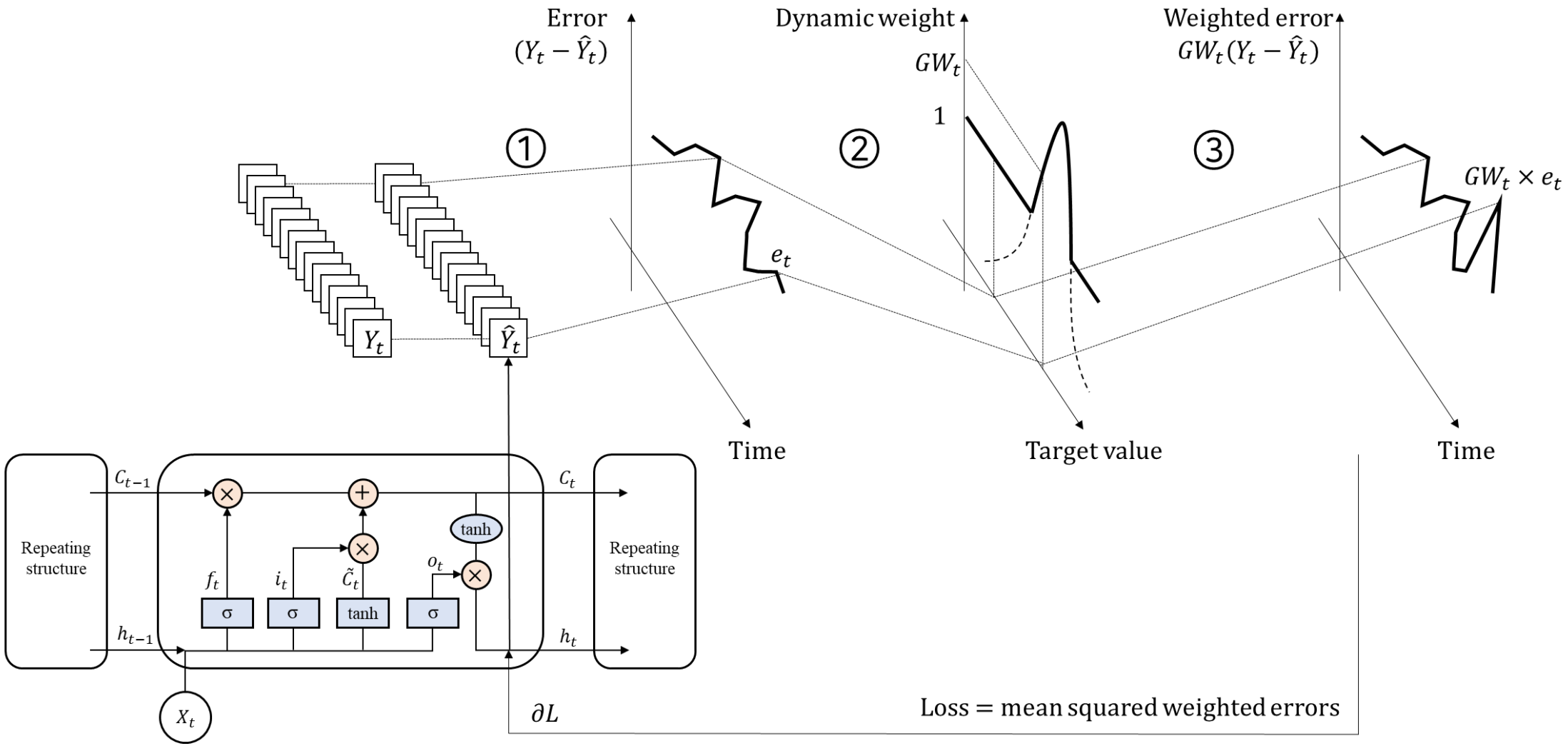
$$\text{MSE} = \sum_{t=1}^{T} \left(Y_t - \hat{Y}_t\right)^2, Y_t \; is \; from \; SGM \text{ and } \hat{Y}_t \; is \; output \; from \; LSTM \; model$$

$$\text{Gaussian weighted MSE} = \sum_{t=1}^{T} GW_t \times \left(Y_t - \hat{Y}_t\right)^2$$

$$GW_t = \begin{cases} a\mathrm{e}^{\frac{-(Y_t - \mathrm{b})^2}{2c^2}}, & |Y_t - b| < c\sqrt{2\ln a} \\ 1, & |Y_t - b| \geq c\sqrt{2\ln a} \end{cases}$$



Methodology – Gaussian weighted loss function



Methodology – LSTM 모델 hyperparameter 최적화

Window size	Epoch	Layer	Hidden neuron	Learning rate	Dropout
12	1500	2	10	0.001	0.3

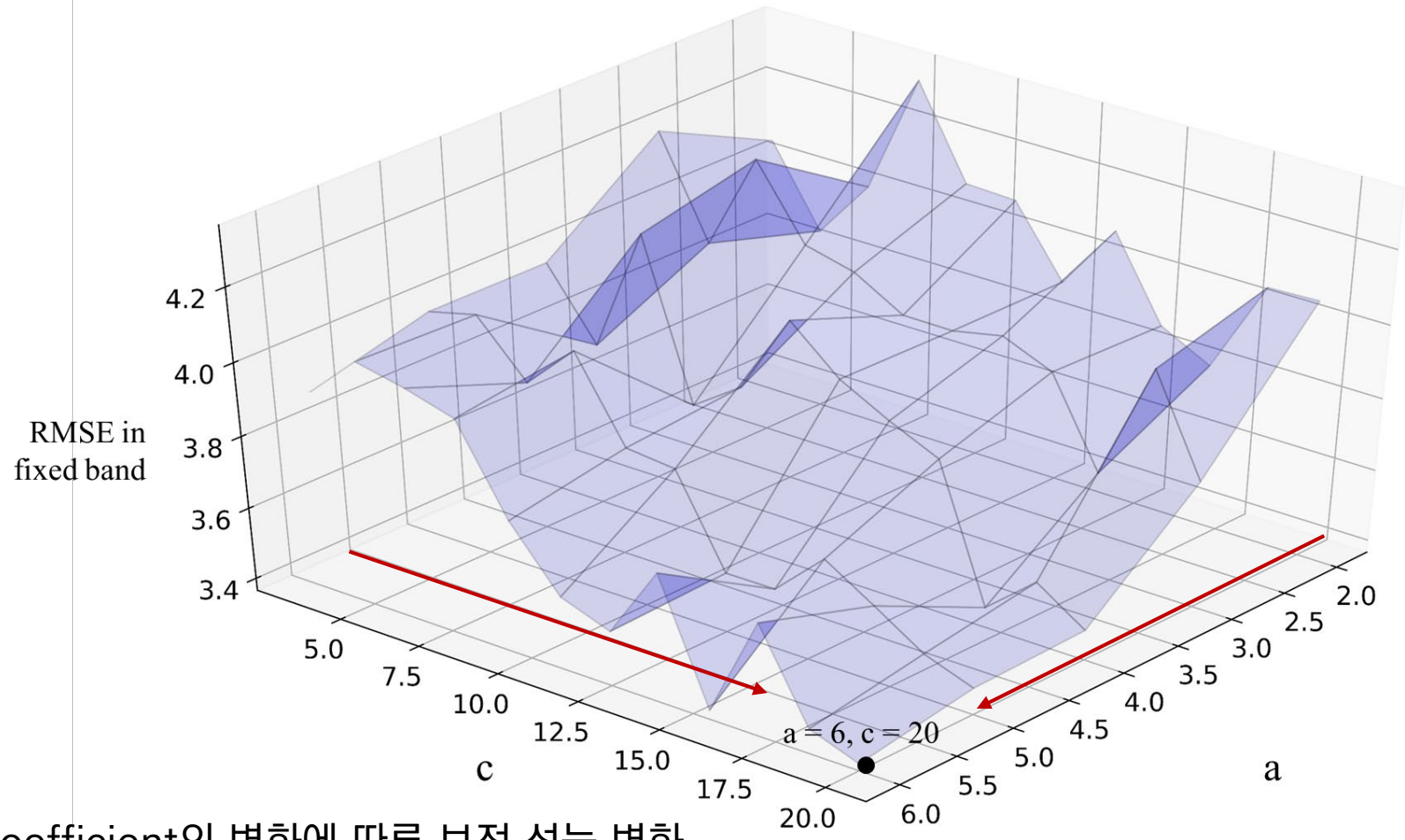
- 학습 조기종료 및 trial and error를 통해 hyperparameter를 최적화하여 모든 모델의 loss function이 global minimum에 도달한 상태에서 성능을 비교함
- Optimizer: Adam
- 성능평가지표: $\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \hat{Y}_t\right)^2}$, $\mathrm{MAPE} = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|$

Result and discussion – LSTM 모델 성능 변화

- Gaussian weighted MSE의 도입으로 보정 능력이 향상된 것을 확인 가능
- Gaussian coefficient가 (6, 20, 20)일 때 가장 좋은 성능을 확인함.

The type of loss function	a	b	c	RMSE	RMSE in fixed band	MAPE in fixed band
MSE	-	-	-	3.50 (0.19)	4.25 (0.23)	17.84 (1.50)
GW[1]	6	20	3.5	3.16 (0.16)	4.057 (0.274)	16.80 (1.15)
GW	6	20	5	3.12 (0.15)	3.98 (0.21)	16.57 (0.91)
GW	2	20	8	3.36 (0.33)	4.10 (0.43)	17.16 (1.96)
GW	6	20	8	3.08 (0.17)	3.96 (0.21)	16.37 (0.89)
GW	6	20	20	3.06 (0.11)	3.77 (0.17)	15.42 (0.68)
GW	2	20	20	3.39 (0.34)	4.25 (0.46)	17.73 (2.05)

Result and discussion – Gaussian coefficient 최적화



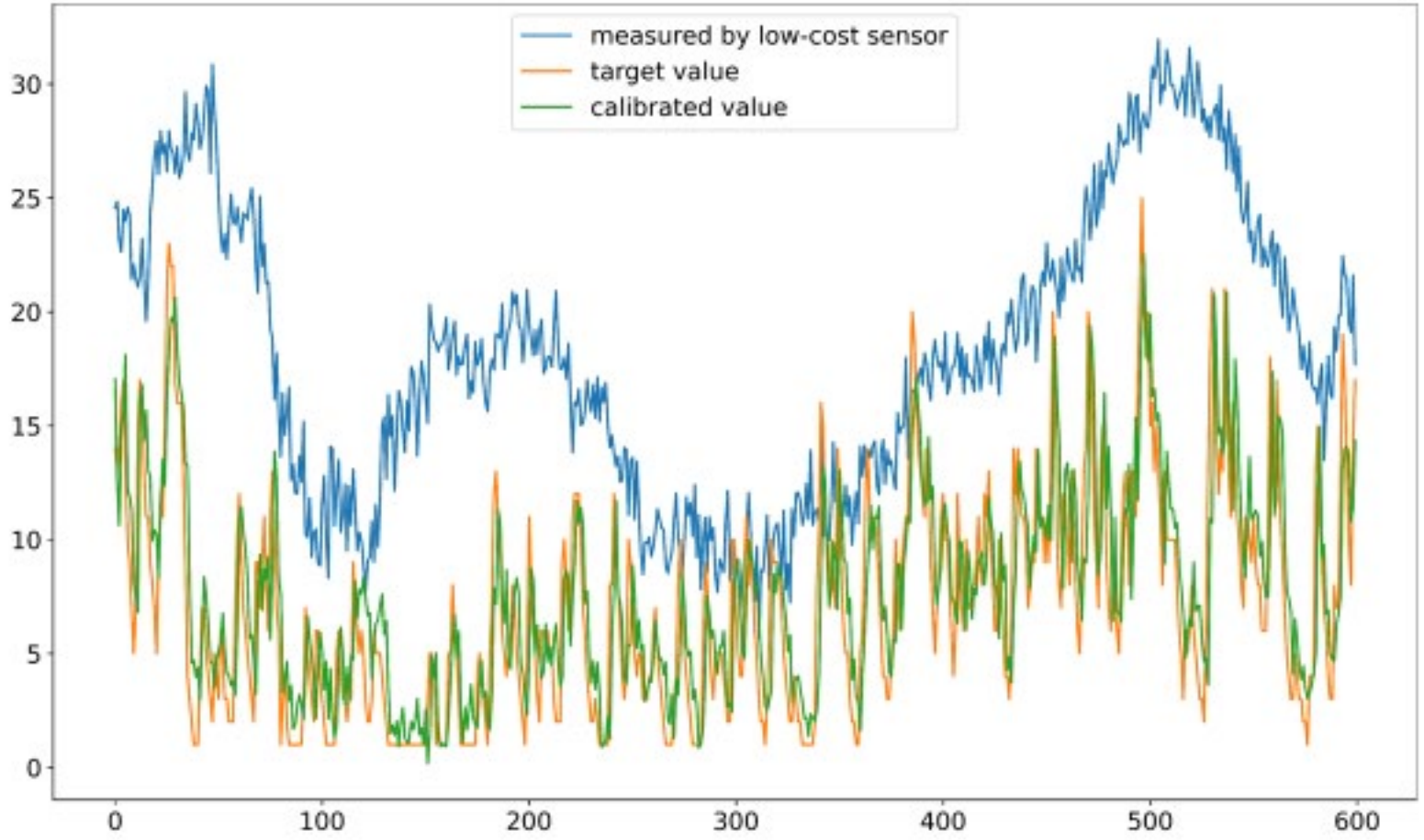
- Gaussian coefficient의 변화에 따른 보정 성능 변화
- Band height(a)가 고정되어 있을 때, band width(c)가 넓어질 수록 보정 능력이 향상됨.
- Band width가(c)가 고정되어 있을 때, band height(a)가 커질 수록 보정 능력이 향상되며, 이 효과는 band height(a)가 커질 수록 더 크게 나타남.
- 보정 성능 향상을 원하는 구간 너머로 길게 dynamic weight를 적용하는 것이 타겟 구간의 성능 향상에 더 유효함

Result and discussion – 보정 전후 비교

베타선 흡수 방식

적외선 산란 방식

보정된 적외선 산란 방식

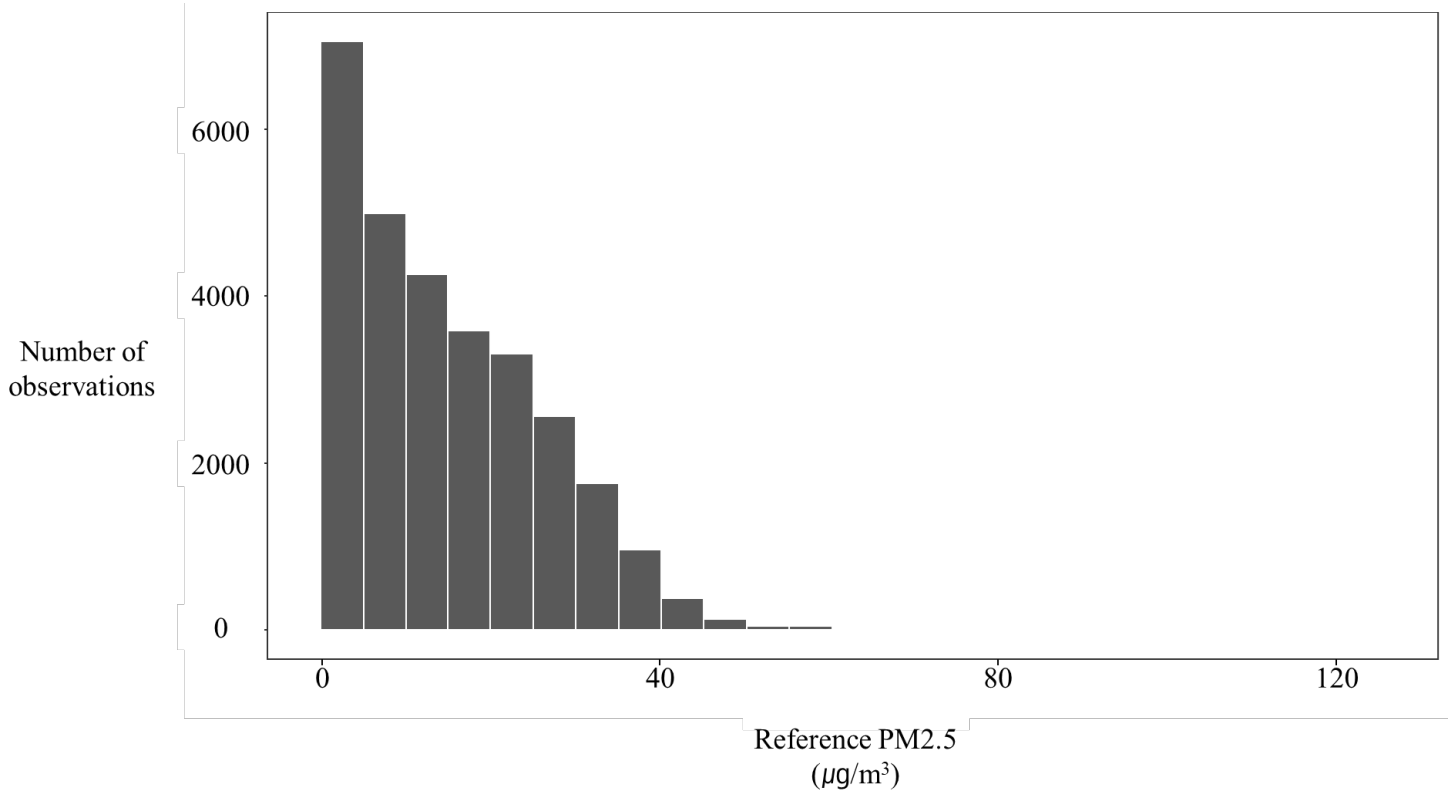


Conclusion – 결론

- LSTM 모델에 주로 사용되는 손실 함수 MSE를 Gaussian weighted MSE로 대체하여 구간에 따른 선택적 보정 능력 향상이 가능함.
- 전체 데이터에 대한 보정 성능의 경우, RMSE를 기준으로 약 12.57%의 성능 향상.
- Band 내부의 보정 성능의 경우, RMSE를 기준으로 약 11.29%, MAPE를 기준으로 약 6.49%의 성능 향상.
- 적외선 산란 방식의 저가형 미세먼지 센서의 성능 향상에 사용 가능함.

Conclusion – 추후 연구

- Extremely right-skewed samples
- 데이터 분포에 따른 LSTM 모델의 학습 및 보정 결과를 비교하기 위해서 추가 데이터 수집 중
- 국내 미세먼지 등급의 기준 농도에 따른 band를 설정하여 기준 농도 근처의 보정 능력을 선택적으로 향상 가능할 것으로 기대됨



Level	Good	Normal	Bad	Very bad
Concentration of PM2.5(μg/m³)	0~15	16~35	36~75	76~