<div align="center">

|  |  |  |
| --- | --- | --- |
| **2019** | **Report – 2019/ 4월3주** | **Seong-gu** |

</div>

**I. Plan for this week**

***Personal Research Subject***

* GCN-LSTM 모델링 개선 방법 계속 진행

**II. Plan for next week**

* 최종 모델링 완성할 것.

**III. Response on Prof. Shin's feedback last week**

Feature를 잘 선택 (데이터 프로세싱으로 정확도 개선)

* 지역은 한국을 전범위로 하지 말고 적당한 수준으로 결정함.
* 효율성 및 정확도 면에서 트레이드 오프 분석 설계 방안을 제기

JEET에 내는 것을 목표로 함.

**IV. Milestone**

* 시공간 데이터 보간 기법 활용 (forward + IDW)
* 추가 feature에 대한 성능 개선 효과 확인 (location predictor, weather, china PM ..)
* feature extraction 방법
* : forward-fixed 기법과 IDW 기법에 대해 가중치를
* . GCN-LSTM 또는 ConvLSTM내부 연산 방법 개선

**V. Reports**

* 개인연구 진행사항

**1. 실험 진행**

**1.1. 데이터 선정**

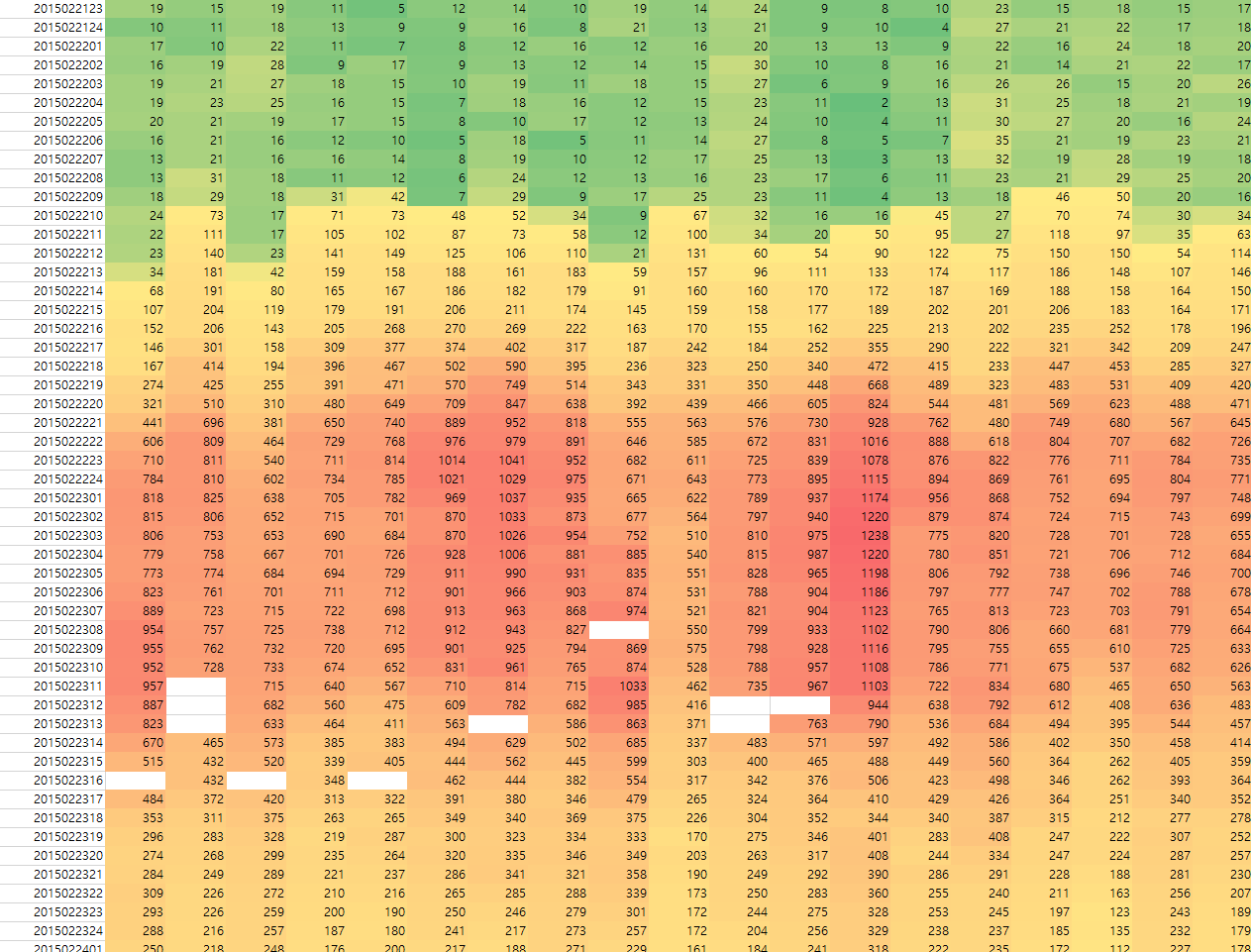
-최적 연구결과를 얻기 위해 training dataset을 늘려 데이터셋을 최소 5개년 이상을 사용하기로 하였습니다. (overfitting 방지), 성능평가에 신뢰성을 늘리기 위해 validation set를 추가하였습니다.

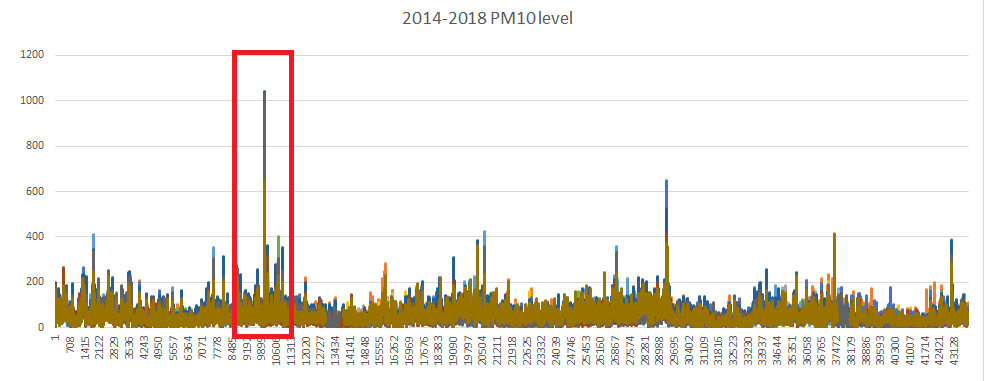
기존 (3개년 데이터 = 2개년 training + 1개년 test)

> 변경 (5개년 이상 데이터 1개년 validation + 3-4개년 training + 1개년 test)

기존 training data에 outliner가 있는 것을 확인하였고 (근 5년간 데이터에서 미세먼지 농도가 1000이 넘는 구간은 해당 구간에서 딱 한 번 존재합니다.) outliner에 맞춰 트레이닝 되는 것을 방지하기 위해 해당 구간은 validation set에 포함시켰습니다.

아래는 해당 구간입니다.





**1.2. 전처리**

1.2.1. 상관관계가 강한 강수량+적설량 Feature를 추가하였습니다.

1.2.2.시간데이터를 One-hot-encoding하여 범주형으로 인식하도록 하였습니다.

1.2.3. 방향을 고려하여 바람 데이터를 보간하였습니다.

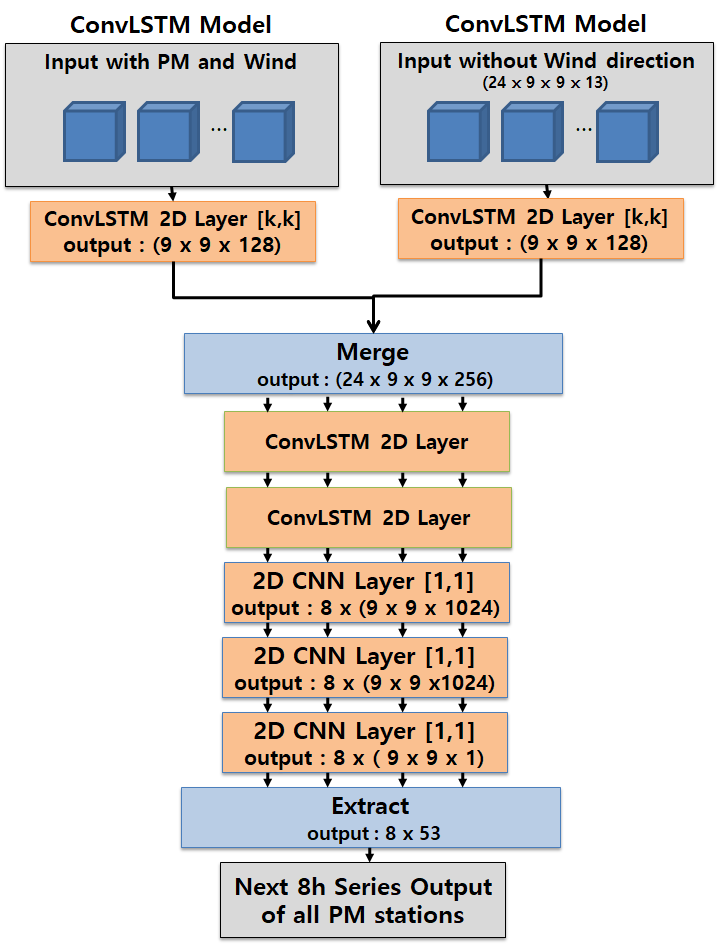
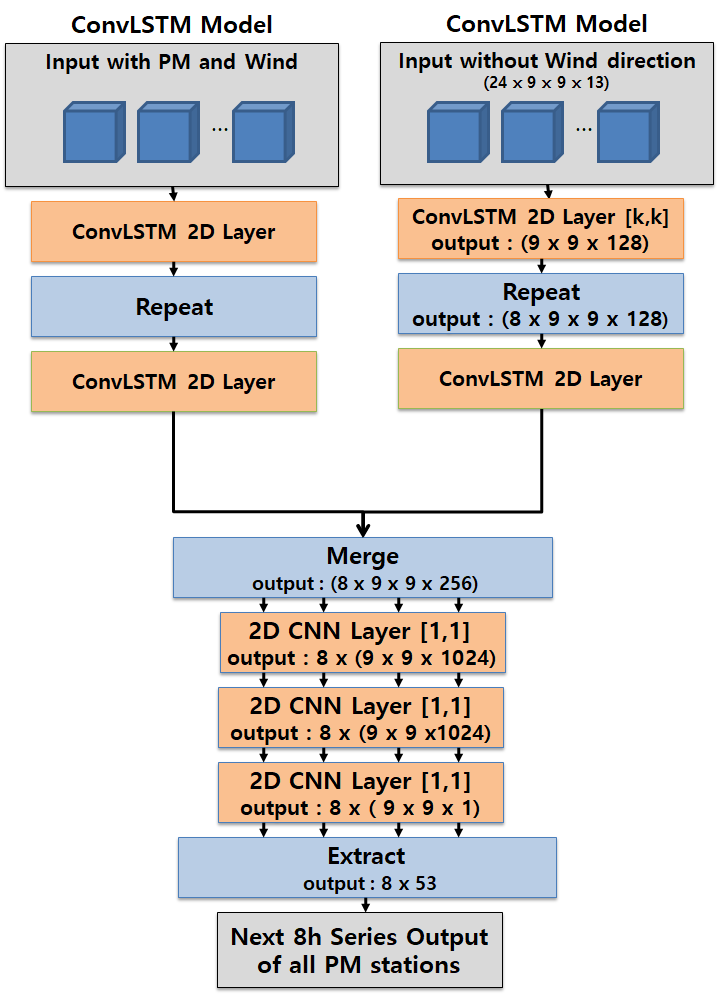
**1.3. 성능평가**

1.3.1. 모델링 개선에 따른 성능 평가

성능 평가할 항목이 많기 때문에 시간 절약을 위해 small scale(9x9 데이터, training data 1개년 치, 테스트 데이터 6개월치)로 진행하였으며 성능은 더 개선될 여지가 있습니다.

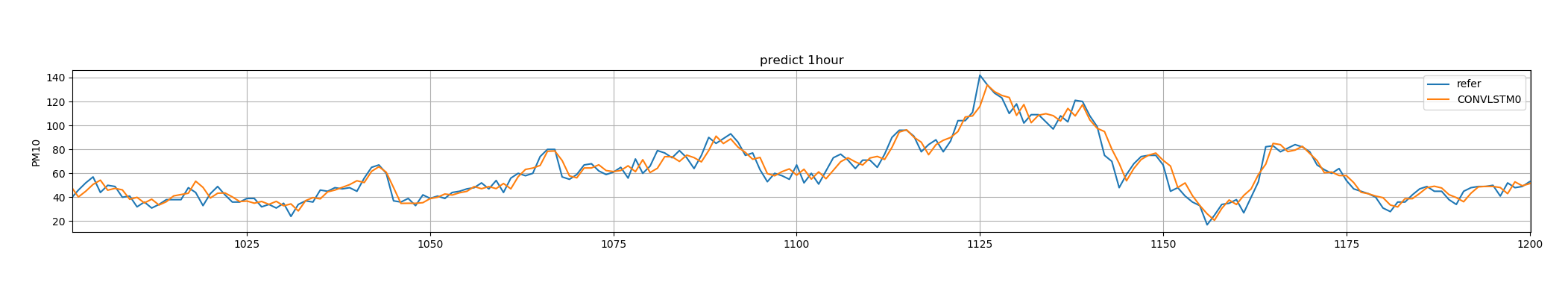
| Model | Prediction hour |
| --- | --- |
| 1h |
| **Follow** | 10.16 |
| **Separate ConvLSTM 3x3  (original)** | **10.00** |
| **Concatenate Double GCNLSTM** | 8.91 |

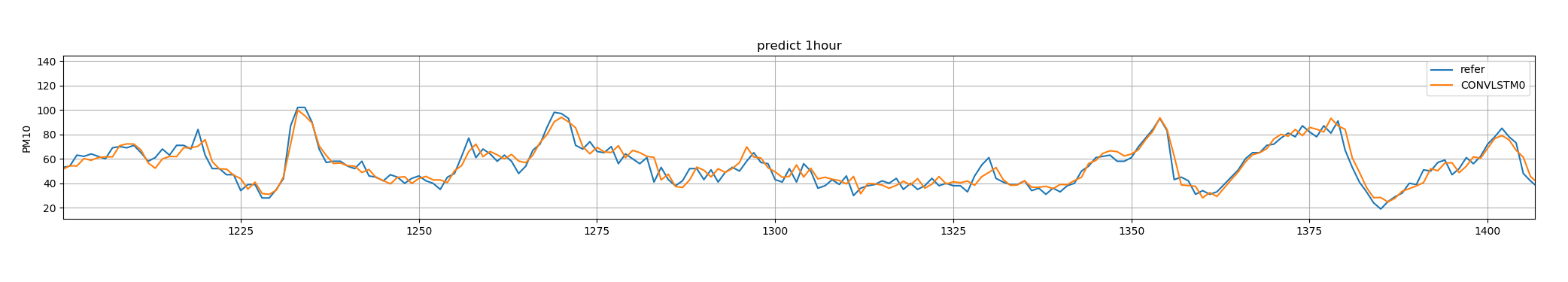
WIND효과와 내부 요인 효과의 input sequence를 decoder 이전에 병합한 모델 (Concatenate Double GCNLSTM) 이 더 좋은 성능을 보여주었습니다. 따라서 해당 모델로 다른 테스트를 진행하였습니다.



변경 전 모델(왼쪽)과 변경 후 모델 (오른쪽)

변경된 모델의 실제값-추정값 비교 그래프입니다. 피크값을 꽤 잘 따라가는 것을 볼 수 있습니다.





1.3.2. 라벨링 데이터에 따른 성능 평가

라벨링 데이터를 실제값의 잔차로 할 때와, 그대로 쓸 때의 결과 차이입니다. 실험조건은 1.3.1과 같습니다.

잔차를 라벨링할 때, 약간의 성능 개선이 있으나, 매우 미미하며 optimal value까지 도달하는 데 필요한 epoch값은 Original signal로 하였을 때 오히려 빨랐으므로 Original signal로 진행하되, 추후 모델이 완성된 다음 잔차를 활용한 개선 방법을 같이 검토할 예정입니다.

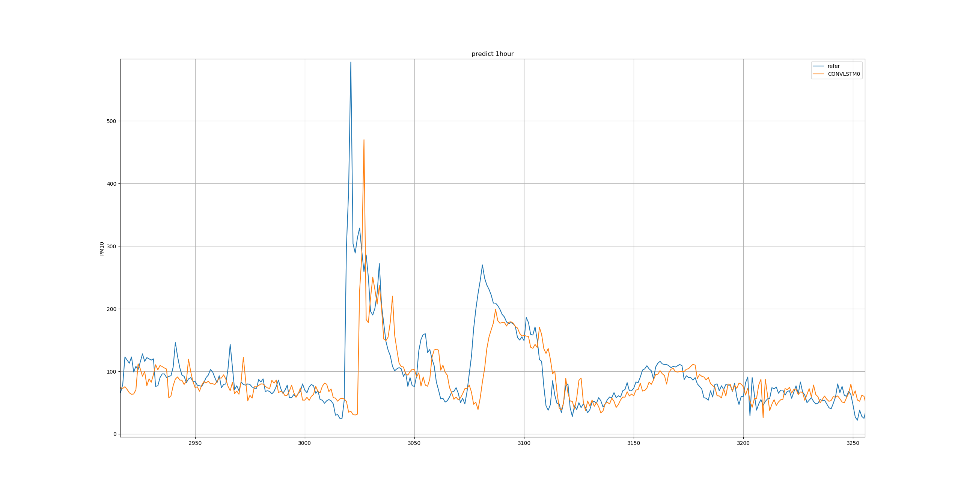
| Model | Prediction hour |
| --- | --- |
| 1h |
| **Differential signal** | 8.91 |
| **Original signal** | 8.97 |

1.3.3. Multi-output 처리 여부에 따른 성능 평가

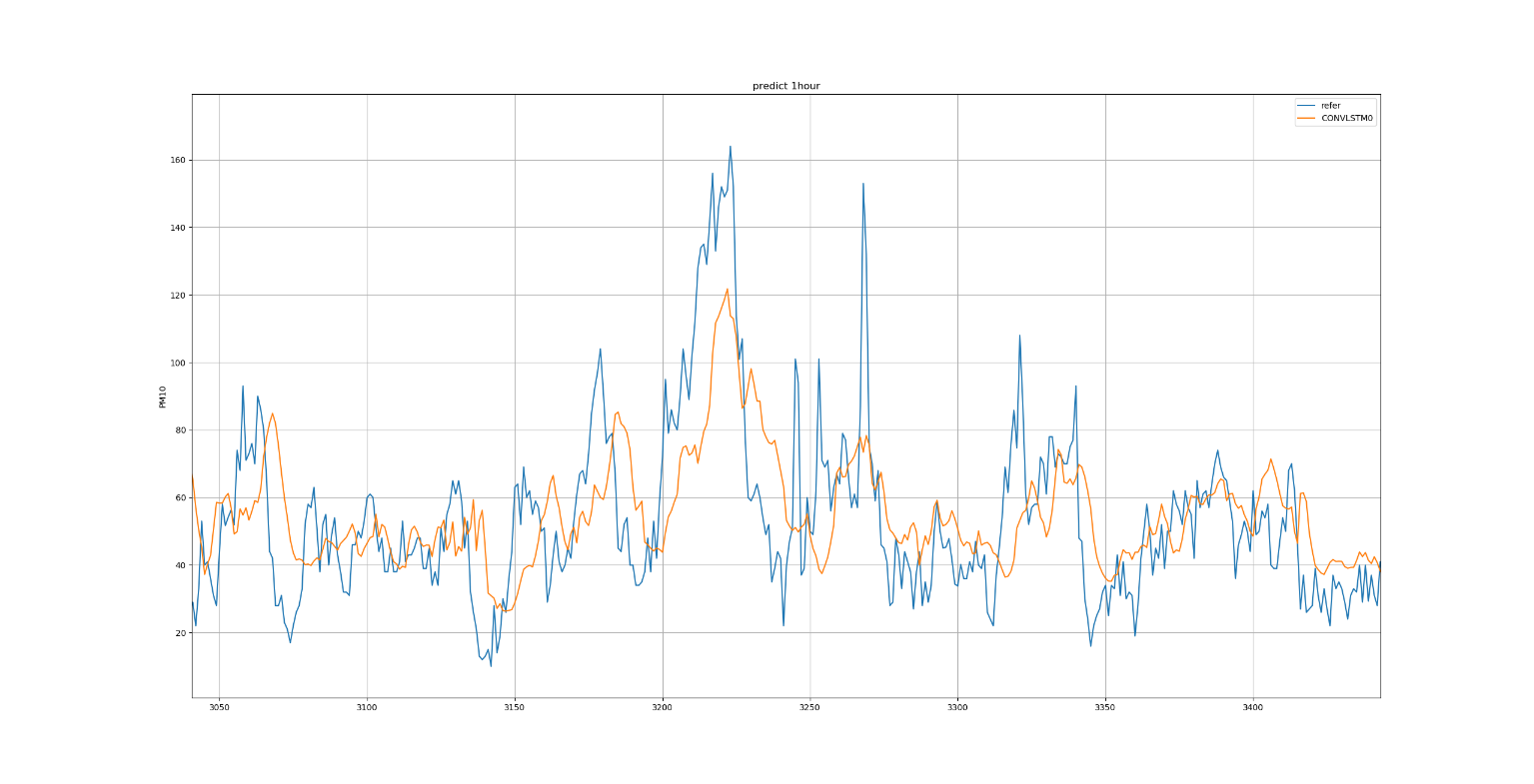
1.3.2의 조건에서 모든 시간대를 한번에 예측하는 multi-step 방법과 원하는 시간대 한번만 예측하는 one-step방법을 비교했습니다. 결과는 제일 예측 시간 term이 긴 시간대의 예측은 성능이 동일하나, 짧은 시간대이면 one-step이 더 나은 결과를 보여줍니다. 이는 loss를 줄이는 학습과정에서 가장 loss가 큰 term이 긴 시간대에 최대한 맞춰서 inner 파라미터가 업데이트되기 때문으로 보입니다. 따라서 각 시간대의 예측은 one-step방식을 쓰는 것으로 정하였습니다.

| Model | Prediction hour | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| 1h | 2h | 3h | 4h | 5h | 6h |
| **Follow** | 10.16 | 14.69 | 17.88 | 20.17 | 21.97 | 23.41 |
| **Multi-step** | 10.09 | **14.31** | **17.09** | **19.01** | **20.49** | 21.66 |
| **One-step** | 8.91 | - | - | - | - | 21.67 |

단, 6h에서의 결과를 그래프에서 확인하면 학습이 잘 안된 것을 볼 수 있습니다. 거의 moving average 효과를 보이고 있는데 이는 training set을 1개년만 사용했기 때문으로 보입니다.



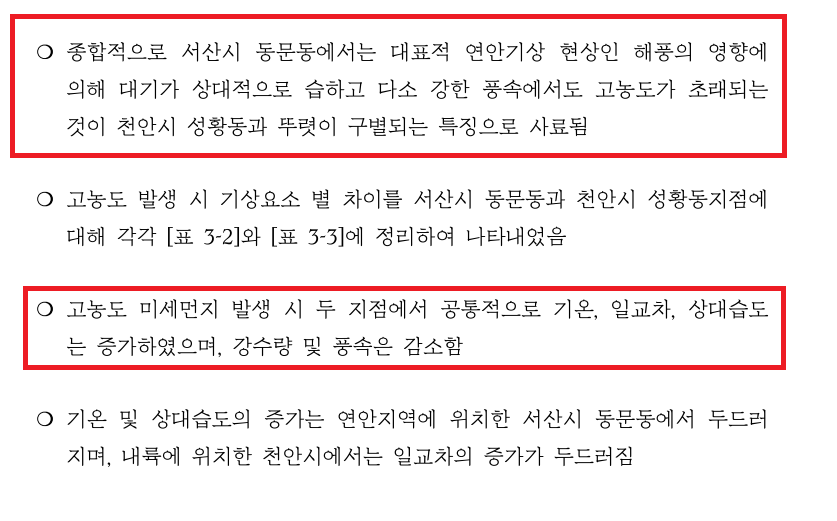
아래는 Training set을 3배로 늘렸을 때의 그래프입니다. (단, validation set은 다른 데이터set을 사용했습니다.) 앞의 moving average효과는 개선되어 보이나, 정확도가 더 개선될 필요가 있어 보입니다.

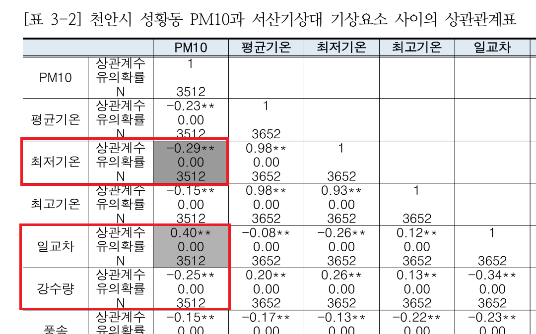
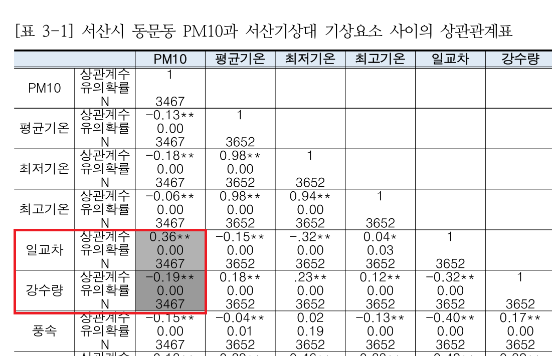


**2. 앞으로의 추가 작업**

**2.1 추가 feature 및 feature extraction**

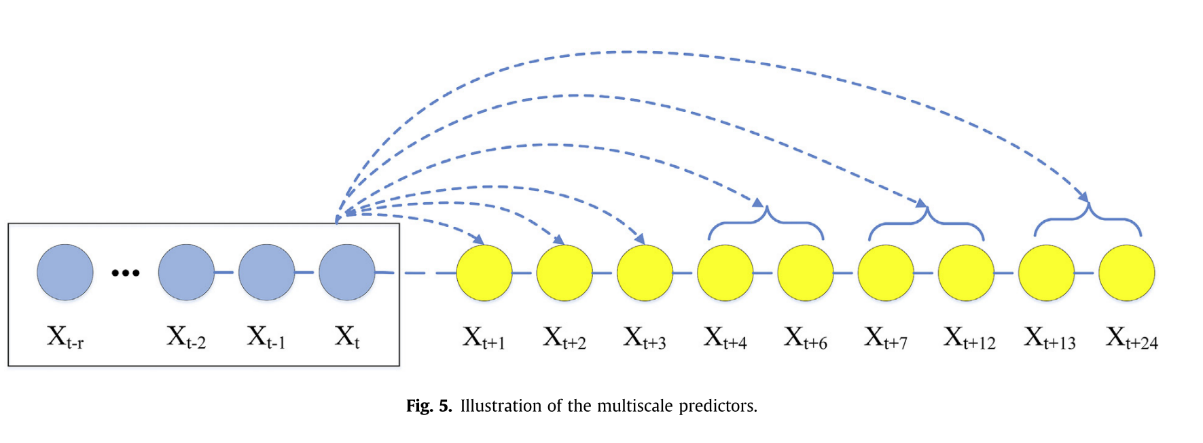
다른 연구 논문을 참고하였을 때, 미세먼지와 상관관계가 높은 주요한 feature들이 있으며, 이 feature에 가중치를 더 주는 방법을 고려할 필요가 있어 보입니다. 또한 측정 환경에 따른 feature의 extraction이 필요하여 이를 뽑아내는 layer를 추가하는 작업을 진행중입니다.

****

****

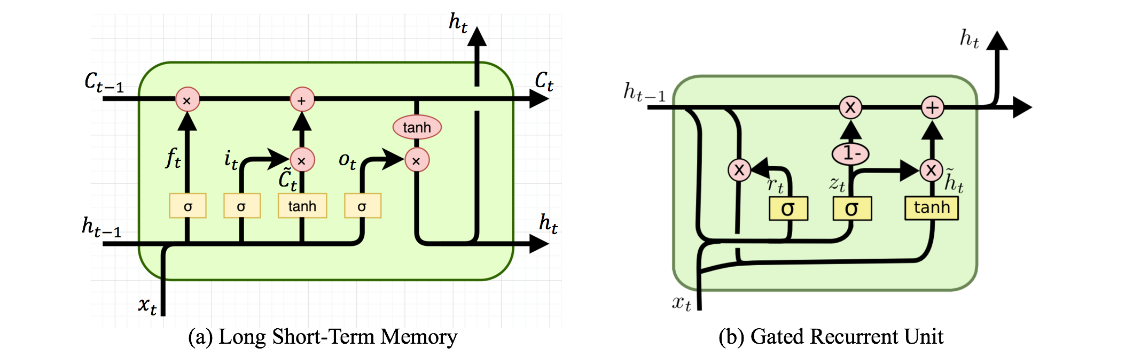
**2.2. 라벨 데이터 변경**

현재까지 결과를 종합할 때, 4h이후에는 미세먼지를 한시간 단위로 정확히 예측하는 것이 상당히 어려운 문제로 보이며 따라서 4h 이후로는 3시간단위 또는 6시간 단위를 잘라 평균을 내고 그 평균값을 예측하는 방향으로 진행하려고 합니다. 실제로 다른 논문에서도 이런 방식으로 결과를 낸 경우가 있습니다.

****

**2.3 모델 개선.**

현재는 LSTM기반 ConvLSTM 또는 GraphLSTM을 사용하고 있으나, 연산속도를 개선하고 파라미터를 줄인 GRU기반 모델링을 사용할 예정입니다. 현재 모델링이 이미 완료되었으며 성능 평가를 진행하고 차용할 예정입니다.



-이외에 Graph Convolutional의 내부 구조를 개선하여 ConvLSTM보다 더 나은 성능을 내기 위한 작업을 진행중에 있으며, 의미있는 결과가 도출되면 업데이트할 예정입니다.

**2.4. 모델링 개선 이외의 성능 개선 방법 평가**

모델 성능이 어느 정도 나와서 모델링을 더 개선할 필요가 없어지면, 다음의 방법들을 시도해볼 예정이며, 성능 향상이 실제로 일어나면 논문에 추가할 예정입니다.

- 데이터 보간 기법에 따른 성능 향상 평가 (시공간 데이터 보간 기법 활용 (forward + IDW)

- china PM을 데이터 feature로 추가했을 때, 장시간 예측에 대한 성능 향상 평가

- input time-step 변화에 따른 성능 평가

- 새로 들어온 데이터 추가 학습에 의한 성능 향상 평가 (weight 불러와서 학습할 때 loss가 생각만큼 안 나오는 특징이 있는데, 다른 파라미터도 같이 넣어줘야 하나봄)

- Decoder있을 때와 없을 때 구분.

- L1L2 정규화를 통한 모델 일반화 성능 향상 평가