2022년 K-water 대국민 빅데이터 공모전 수행 결과보고서

제 목	하수처리장 수질 예측모델 개발			
부 문	데이터 융합	●	제품 및 서비스 개발	○
성 명	팀 장	김서린	010-6439-6692	
		중앙대학교	tjfls96@naver.com	
	팀 원	김소은	중앙대학교	
		박상우	중앙대학교	
		허인	중앙대학교	

Ⅰ. 과제 목표

하수처리장은 생활하수를 대상으로 처리하는 시설인 만큼 현재 하수도 수질관리를 위한 여러 관리 방안이 요구되고 있다. 하지만 하수도 시설의 전반적 운영이 행정구역을 중심으로 이루어짐에 따라 효율적 관리가 어려운 상황이다. 최적의 수질관리 시스템을 개발하여 광역적으로 적용한다면 효율적인 관리가 가능해질 것이다. 수질예측을 통해 녹조 발생에 선제 대응하고 수질사고를 미연에 방지하여 안전한 물을 공급할 수 있다. 따라서 본 연구는 하수처리장의 수질관리 선진화를 위하여 하수처리장의 수질을 예측하는 최적의 모델을 찾고자 한다.

Ⅱ. 활용 데이터

한국수자원공사에서 제공하는 ‘하수처리시설 방류 수질 현황(일자료)’의 OpenAPI 데이터를 사용하였다.

하수처리장 수질 예측의 목적은 궁극적으로 유출 수질을 예측하는 목적과 동일하다고 판단하여 방류수 데이터를 예측하였다. 선행연구를 조사하는 과정에서 수질 예측을 위해 회귀분석을 진행할 때 COD, TN을 종속변수로 채택하는 것을 확인하였다. COD는 화학적

산소요구량으로, 미생물이 분해하지 못하는 유기물을 화학적 산화제가 직접 산화시키는데 소모된 양에 따른 전자 이동량을 산소 필요량으로 환산한 값이다. TN은 수중에 포함된 무기질 질소 및 유기성 질소의 질소량 합계로, 질소가 많을수록 녹조의 생육을 촉진시킨다. 따라서 COD, TN이 높을수록 수질 오염이 심하다는 의미이다. 이를 바탕으로 방류수 COD, 방류수 TN을 최종 예측할 변수로 선정하였다.

최종적으로 결측치가 없었던 '장항 공공 하수처리시설'의 2019년 1월 1일부터 2021년 10월 31일까지의 데이터를 사용하기로 결정하였다.

다음은 예측 모델을 설계하기 위해 정리한 데이터의 형태이다. 시계열 데이터이기 때문에 시간 순서에 따라 방류수 COD, 방류수 TN을 정리하였다.

	wqdt	bCod	bTn
0	2019-01-01	14.4	10.370
1	2019-01-02	14.0	10.181
2	2019-01-03	12.2	9.746
3	2019-01-04	12.6	8.880
4	2019-01-05	15.0	9.691
...	...	...	...
1030	2021-10-27	7.5	6.286
1031	2021-10-28	7.6	5.268
1032	2021-10-29	6.7	5.438
1033	2021-10-30	6.9	5.834
1034	2021-10-31	7.2	6.370

1035 rows × 3 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1035 entries, 0 to 1034
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype         
---  ------  --------------  -----         
 0   wqdt    1035 non-null   datetime64[ns]
 1   bCod    1035 non-null   float64       
 2   bTn     1035 non-null   float64       
dtypes: datetime64[ns](1), float64(2)
memory usage: 24.4 KB
```

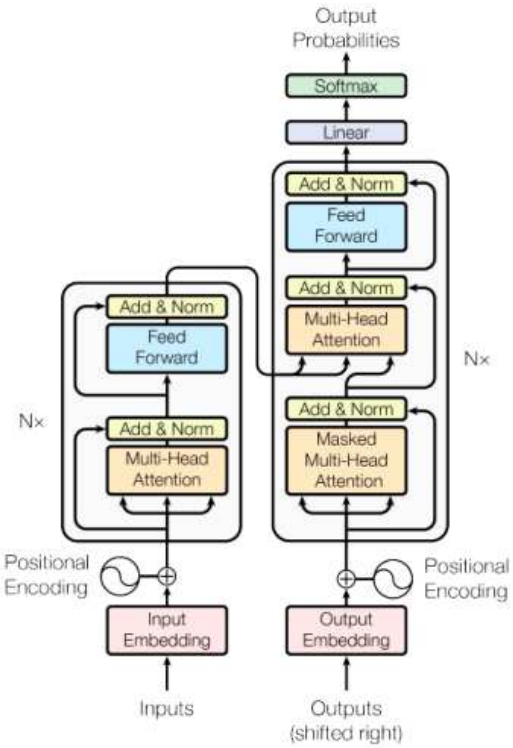
Ⅲ. 주요 내용

예측 모델은 베이지안 LSTM, Transformer, GRU 총 3가지를 사용하여 방류수 COD, 방류수 TN을 예측하였다. MSE를 통해 가장 성능이 좋은 모델을 선정하고자 한다.

1. 이론적 배경

가. Transformer

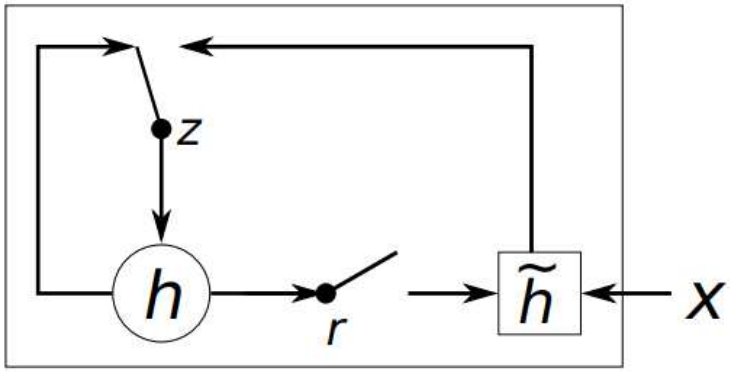
Sequential 문제를 풀기 위해 입력값과 출력값 sequence를 활용하는 순환구조의 모델을 사용했다. 그러나 병렬화가 제한되고, sequence의 길이가 길어질수록 취약해지는 문제가 있었다. 이러한 점을 보완하기 위해 Transformer는 인코더-디코더의 구조를 따르면서 attention 메커니즘만을 사용해 구현한 모델이다. Attention 메커니즘을 활용해 입력값의 representation을 계산하여 feature을 추출하는 것이 특징이다.



각각 6개의 인코더와 디코더가 쌓여있는 구조이며 인코더에서 입력 시퀀스를 입력 받고, 디코더에서 출력 시퀀스를 출력한다.

나. GRU(Gated Recurrent Unit)

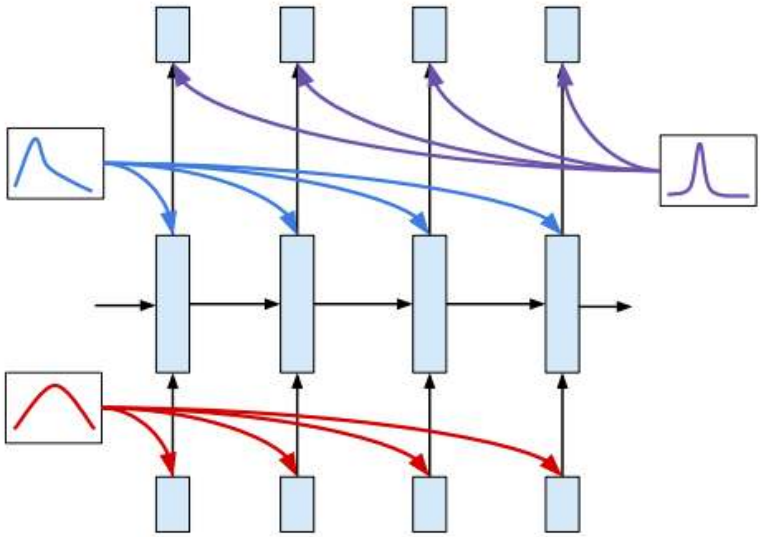
파라미터는 많아지는데 데이터가 적은 경우 과적합이 발생하는 LSTM의 문제를 개선한 모델이다. GRU는 두 개의 gate로 이루어져 있는데, 하나는 LSTM의 forget gate와 input gate를 통합한 하나의 update gate와 reset gate이다. LSTM에 비해 파라미터의 수가 적지만 성능이 비슷한 모델이다.



Reset gate(그림의 r)은 과거의 hidden state의 값(그림의 h)을 버려주는 역할을 한다. Update gate(그림의 z)는 hidden state가 새로운 hidden state로 갱신이 되는지를 결정한 뒤, 현재 유닛의 입력값(그림의 x)과 reset gate의 입력값을 이용해 최종 hidden state(그림의 h˜)를 계산한다.

다. 베이지안 LSTM

정확성을 높이기 위해 local gradient 정보를 대략적인 사후확률 분포에서 샘플링한 정보와 통합하는 구조이다.

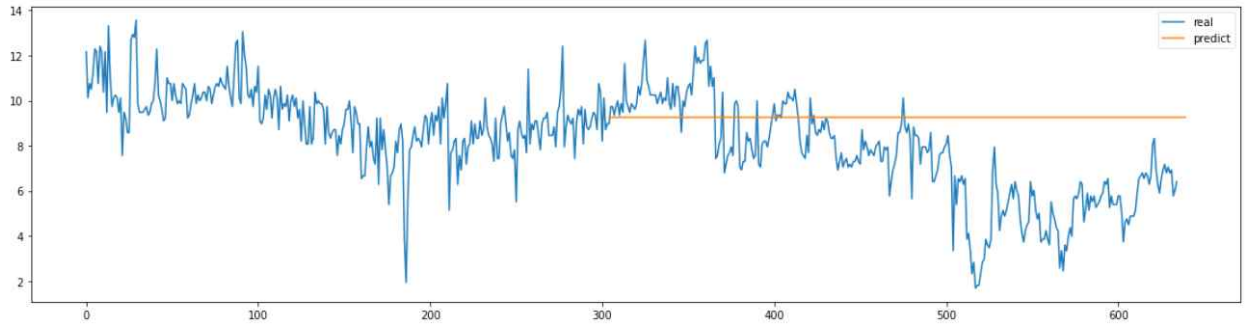


시간적 흐름에 맞게 잘린 역전파의 단순한 적용이 불확실하지만 좋은 성능의 추정치를 산출한다. 학습 시 아주 적은 추가 계산 비용으로 좋은 정규화를 할 수 있는 장점이 있다.

2. 모델 성능 비교 및 결과

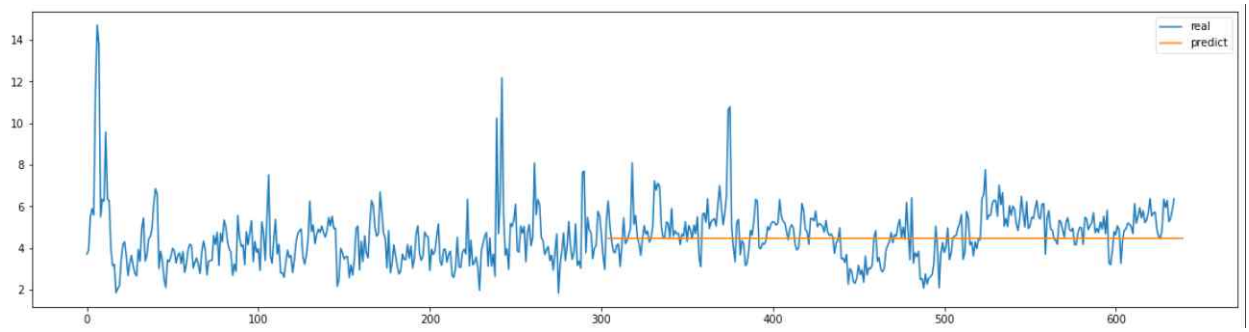
가. Transformer

- 방류수 COD 예측



100 epoch동안 학습했고, MSE 값은 8.4859이다.

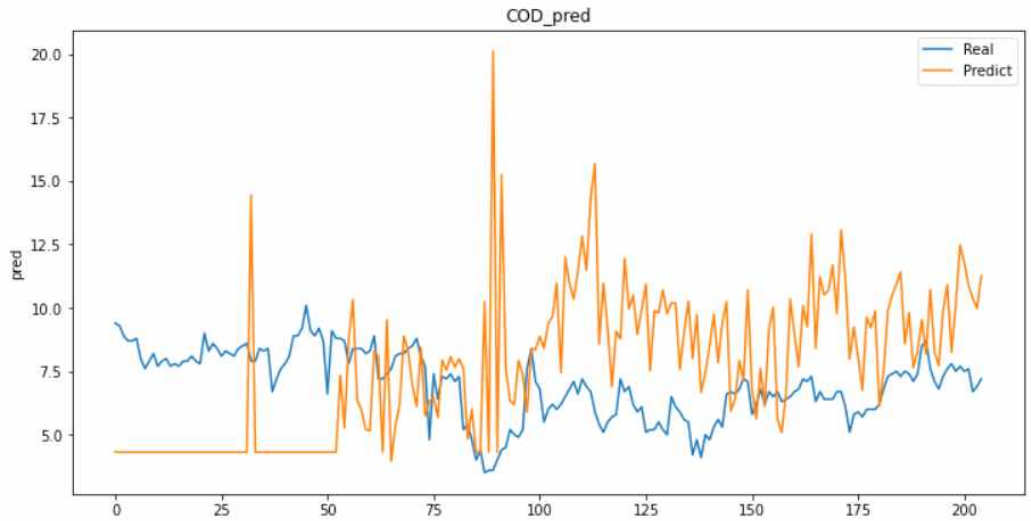
- 방류수 TN 예측



100 epoch동안 학습했고, MSE 값은 1.39935이다.

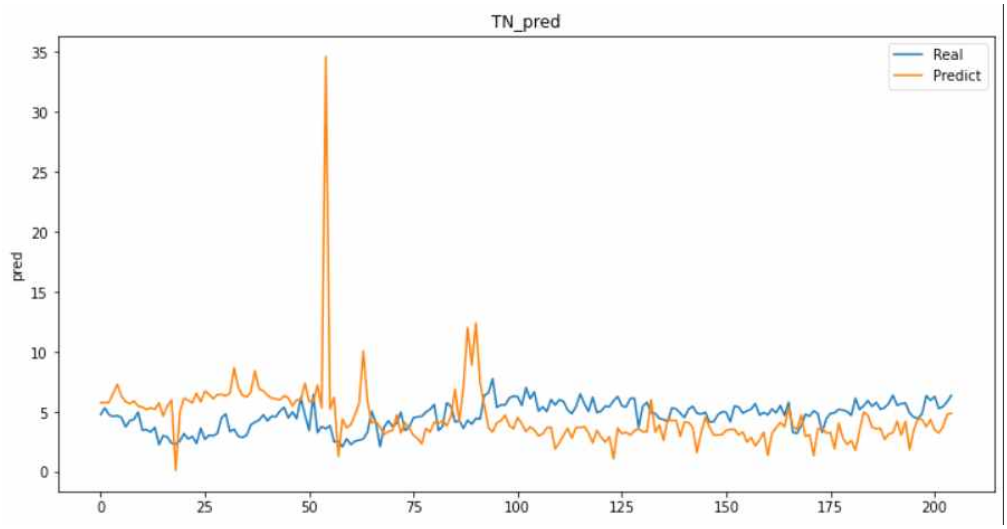
나. GRU

- 방류수 COD 예측



100 epoch동안 학습했고, MSE 값은 13.7470이다.

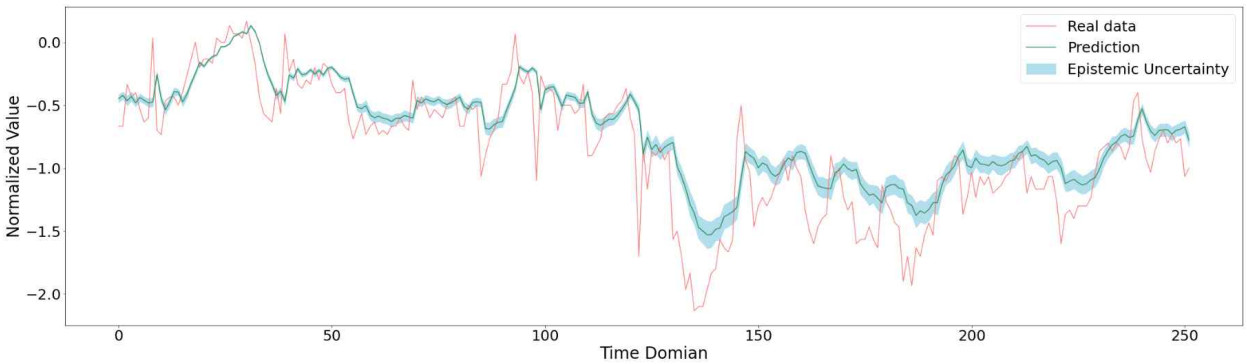
- 방류수 TN 예측



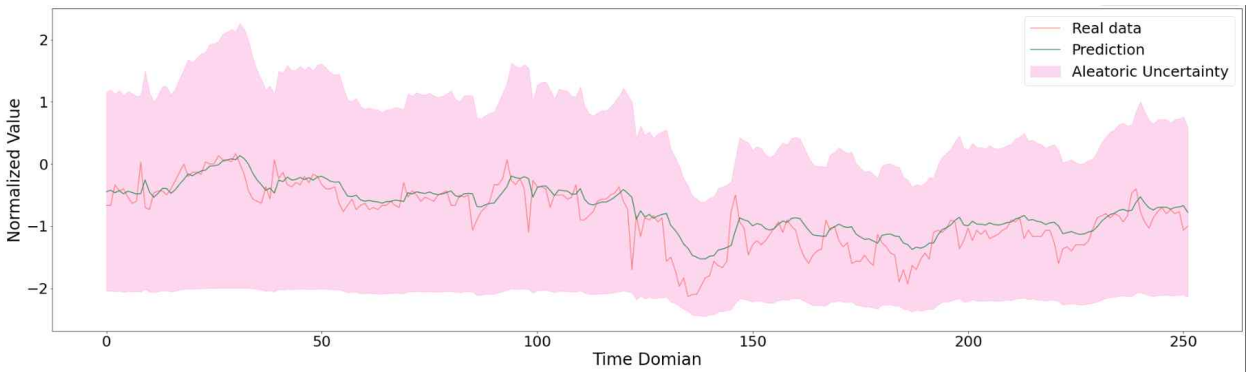
100 epoch동안 학습했고, MSE 값은 10.4618이다.

다. 베이지안 LSTM

- 방류수 COD 예측



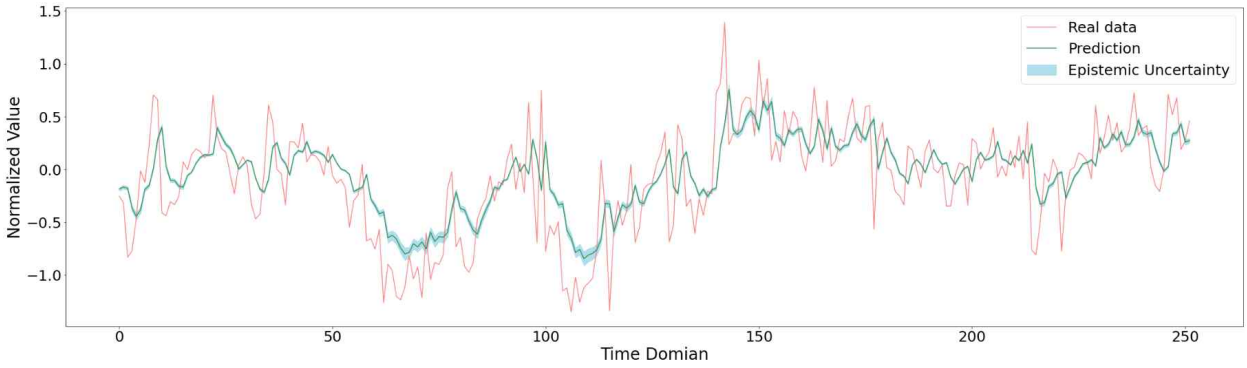
(COD Epistemic Uncertainty)



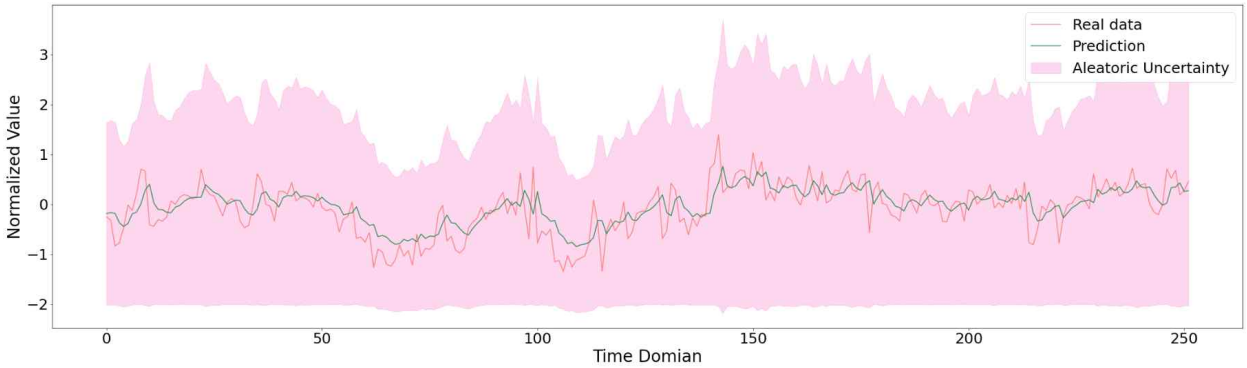
(COD Aleatoric Uncertainty)

학습이 조기 종료되어 38 epoch동안 학습했고, MSE는 0.66이다.

- 방류수 TN 예측



(TN Epistemic Uncertainty)



(TN Aleatoric Uncertainty)

학습이 조기 종료되어 44 epoch동안 학습했고, MSE는 0.54이다.

3. 하이퍼 파라미터 튜닝

최종 모형으로 베이지안 LSTM을 선정하여 하이퍼 파라미터 튜닝(매개변수 검토)을 진행하였다.

LSTM 모형을 구축할 때의 데이터는 장항 하수처리장 수질 데이터로, 2019년 1월 1일부터 2021년 10월 31일 까지의 1035개의 데이터 셋을 활용하였다. 75%에 해당하는 2021년 2월 15일 까지의 776개의 데이터 셋을 훈련, 이후 25%에 해당하는 259개의 데이터 셋을 테스트 셋으로 사용하였다.

LSTM의 과적합을 피하기 위하여 Dropout과 Recurrent dropout을 사용하였다. 이들은 딥 러닝 학습시의 과적합을 해결하기 위해 노드 일부를 중지시키는 학습 기법이다.

콜백 함수로는 Early Stopping Method를 사용하였다. 이는 설정한 Epoch 동안 모델의 성능이 증가하지 않을 때 학습을 중단하는 기법이다. Loss를 관찰해 15개의 epoch동안 모델의 성능이 증가하지 않을 때 자동으로 학습을 멈추는 콜백 함수를 지정하였다.

튜닝할 하이퍼 파라미터는 노드의 개수, 심층 레이어의 개수, Dropout의 비율, Recurrent dropout의 비율이다.

비교를 위한 오차로는 평균 제곱근(MeanSquare Error)을 사용하였고, 최적화 알고리즘으로 Adam optimizer를 사용하여 Epoch을 1000으로 설정해 모델을 학습하였다.

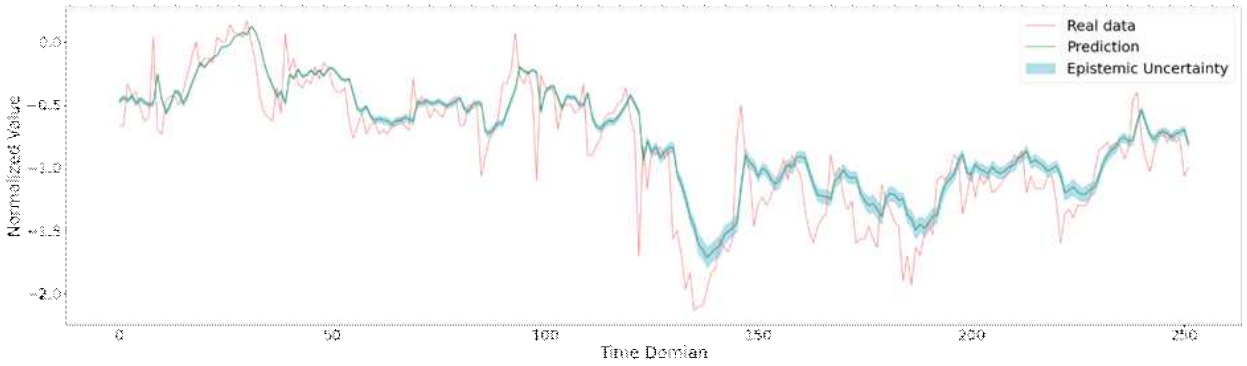
하이퍼파라미터 튜닝 결과 선정된 최종 모델은 각각 다음과 같다.

	COD	TN
노드의 수	512	512
심층 레이어의 수	1	1
Dropout	0.1	0.1
Recurrent dropout	0.2	0.3

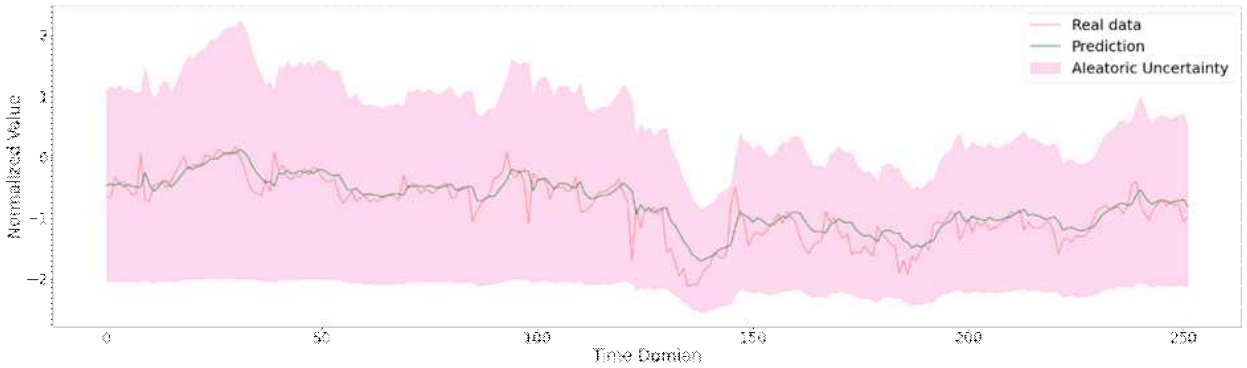
(베이지안 LSTM 모델과 Parameter)

	COD	TN
Mse	0.52	0.54
R2 score	0.76	0.51

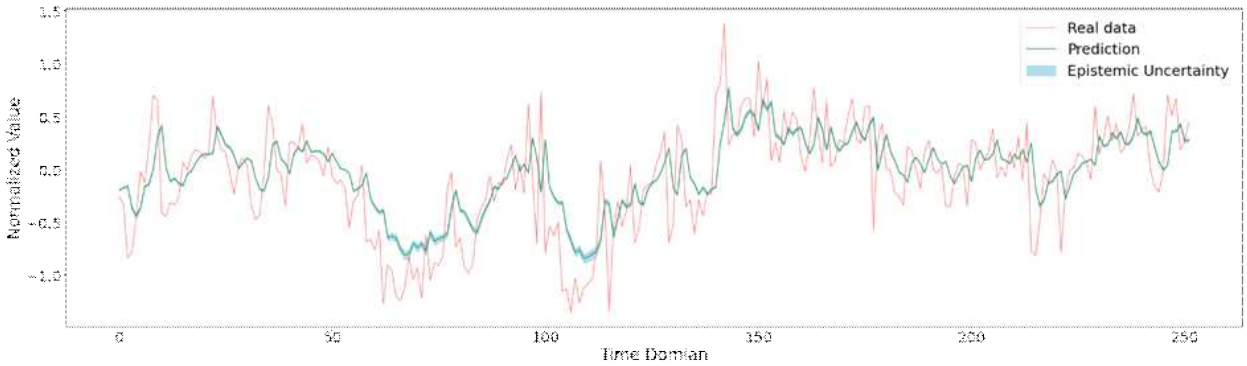
(Mse와 R2 score)



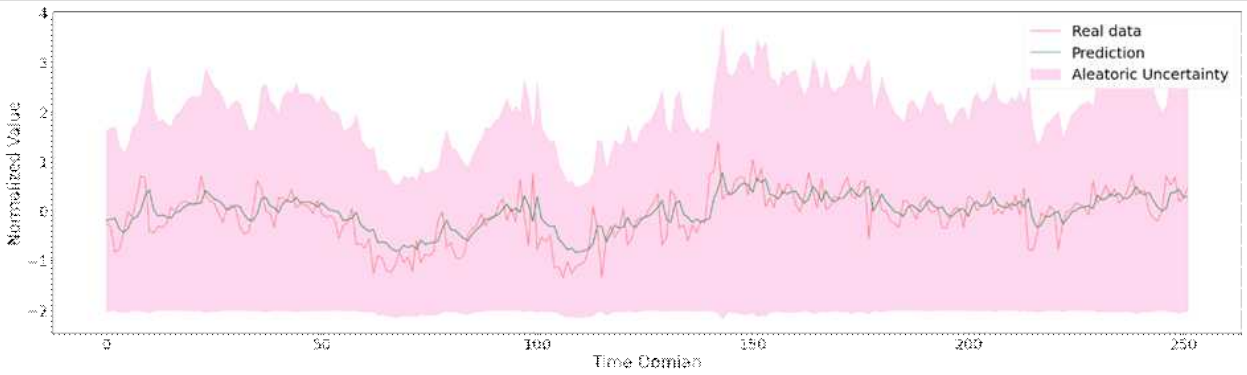
(COD Epistemic Uncertainty)



(Cod Aleatoric Uncertainty)



(Tn Epistemic Uncertainty)



(Tn Aleatoric Uncertainty)

베이지안 LSTM 모형은 aleatoric과 epistemic uncertainty를 개별적으로 모델링한다. Aleatoric uncertainty는 데이터 자체에 담겨 있는 고유 노이즈로 인한 불확실성이고, epistemic uncertainty는 모델이 데이터를 충분히 설명하지 못하는 uncertainty이다.

COD와 TN 모델 모두 aleatoric uncertainty가 높지만 epistemic uncertainty가 낮고 예측치와 관측치 사이의 차이가 적다. 즉 모델이 데이터를 충분히 설명한다고 할 수 있으며, 따라서 베이지안 LSTM 모델의 예측력이 우수하다고 결론지을 수 있다.

Ⅳ. 결과 및 기대효과

하수처리장 수질 예측 모델 개발은 운전자의 경험적 지식에 의존하여 운영함으로써 발생하는 신뢰성, 객관성 등의 문제를 해결할 것으로 보인다.

하수처리장 수질에 영향을 미치는 COD, TN 양의 추세를 파악하고 예측함으로써 수질관리를 위한 구체적인 운전 전략을 마련할 수 있다. COD 및 TN의 예측값이 일정 기준보다 초과될 경우 하수처리장 운전조건에 맞지 않다고 판단하여 운전조건을 재설정할 수 있다. 또한 수질을 예측함으로써 수질 사고가 일어나기 전에 선제 방안을 마련할 수 있을 것이다. 따라서 수질 예측은 환경 변화와 운전 전략에 영향을 받는 하수처리장 운영의 특성상, 매우 필수적인 자료로 활용될 수 있을 것이라 예상한다.