<https://colab.research.google.com/drive/1vwRKZHGpOm2_vt4YLDpAERKJ4BMOyaYl?usp=sharing>

**[요약]**

SAC(Soft Actor-Critic)는 최대 엔트로피 원칙을 도입하여 탐색적인 정책을 학습하고, 오프폴리시 학습을 통해 데이터 효율성을 높인다. 이러한 특징으로 인해 SAC는 복잡한 연속적인 행동 공간에서 강력한 성능을 발휘한다. 특히, 연속적인 값을 가지는 행동 선택에 적합하다.

LSTM(Long Short-Term Memory) 모델을 SAC 알고리즘에 사용하여 내부 온도를 예측하는 방식을 도입하면 시계열 데이터 처리를 강화할 수 있다. LSTM은 시계열 데이터의 시간적 종속성을 학습하는 데 강력한 도구이므로, 이를 SAC와 결합하면 과거의 정보를 반영하여 더 나은 결정을 내릴 수 있다.

모듈 사용 시 수정한 값은 다음과 같다:

learning\_starts, batch\_size, learning\_rate, buffer\_size

행동에 대한 보상과 정책 설정에 주의해야 한다. SAC는 Q값에 민감한 모델이므로, 보상 함수 설정에 유의해야 한다.

LSTM 층을 추가하여 시간의 흐름을 고려하도록 정책을 변경하는 작업은 현재 진행 중이다. 기존의 맞춤형 손실 함수 기반의 머신러닝 모델보다 시간 측면에서 개선되었다. 또한, 개폐도 범위 제한이 적절하게 학습되도록 추가되었다.

**Pytoch 기반의 stable\_baselines3의 SAC(Soft Actor-Critic) 모델 사용한 최적화 모델**

SAC 모델 선택 이유: 연속적인 행동 공간, 안정적 성능

Soft Actor-Critic (SAC) 알고리즘은 강화학습의 한 종류로, 오프폴리시(Off-Policy) 학습과 최대 엔트로피(Maximum Entropy) 원리를 결합하여 높은 탐색 성능과 학습 안정성을 제공함. SAC는 특히 연속적인 행동 공간을 다루는 데 효과적.

**SAC 알고리즘의 주요 특징**

1. 오프폴리시 학습(Off-Policy Learning):

에이전트가 데이터를 수집하는 정책과 학습하는 정책이 다름.

이를 통해 데이터 효율성을 높이고, 이전에 수집된 경험을 재사용할 수 있음.

2. 최대 엔트로피 원리(Maximum Entropy Principle):

정책을 최대 엔트로피 원칙에 따라 학습함. 이는 단순히 보상을 극대화하는 것이 아니라, 높은 엔트로피(즉, 불확실성)를 유지하면서 보상을 극대화하는 것을 의미함.

이로 인해 탐색이 강화되고, 최적 정책을 찾는 과정에서 더 다양한 행동을 시도할 수 있음.

**SAC의 기본 구성 요소**

1. Actor (정책 네트워크):

현재 상태에서 어떤 행동을 취할지 결정함.

확률적 정책을 사용하여 주어진 상태에서 행동을 샘플링함.

2. Critic (가치 네트워크):

주어진 상태에서 특정 행동의 가치를 평가함.

Q-값(특정 상태에서 특정 행동을 취했을 때 얻을 수 있는 예상 보상의 값)을 예측하여 Actor의 행동을 평가하고 업데이트하는 데 사용됨.

3. Entropy Term:

정책의 엔트로피를 계산하여 정책 업데이트 시 사용함.

엔트로피 항을 추가하여 정책이 더 탐색적으로 유지되도록 함.

**SAC 알고리즘의 작동 방식**

1. 초기화:

정책 네트워크(Actor)와 두 개의 Q-함수 네트워크(Critic)를 초기화함.

경험 재생 메모리(Experience Replay Buffer)를 초기화함.

2. 경험 수집:

환경에서 현재 정책에 따라 행동을 취하고 경험을 수집하여 재생 메모리에 저장함.

3. Q-함수 업데이트:

Critic 네트워크를 업데이트함.

손실 함수:

L(Qi​)=E(s,a,r,s′)∼D​[21​(Qi​(s,a)−(r+γ(j=1,2min​Qtarget,j​(s′,a′)−αlogπ(a′∣s′))))2]

Qi​: Critic 네트워크

D: 경험 재생 메모리에서 샘플링된 데이터

γ: 할인율

α: 엔트로피 계수 (탐색과 학습의 균형을 조절)

4. 정책 업데이트:

Actor 네트워크를 업데이트함.

손실 함수:

L(π)=Es∼D,a∼π​[αlog(π(a∣s))−Q(s,a)]

π: 정책 네트워크

Q: Critic 네트워크가 예측한 Q-값

5. Target 네트워크 업데이트:

안정적인 학습을 위해 타겟 네트워크를 천천히 업데이트함.

6. 반복:

위 과정을 반복하여 정책과 Q-함수를 점진적으로 개선함.

**결론**

SAC는 최대 엔트로피 원칙을 도입하여 더욱 탐색적인 정책을 학습하며, 오프폴리시 학습을 통해 데이터 효율성을 높임. 이러한 특징으로 인해 복잡한 연속적인 행동 공간에서 강력한 성능을 발휘함.

이와 같이 연속적인 값을 가지는 행동 선택에 적합함. (개폐도는 연속적인 값임.)

LSTM(Long Short-Term Memory) 모델(내부 온도 예측 모델)을 SAC(Soft Actor-Critic) 알고리즘에 사용**(현재 상태 정보와 행동을 기반으로 LSTM 모델로 다음 내부 온도를 예측하고, 예측한 내부 온도와 현재 행동을 다음 상태 정보에 업데이트해서 반환시킴.)**하면 시계열 데이터 처리를 강화함.

LSTM은 시계열 데이터의 시간적 종속성을 학습하는 데 강력한 도구이므로, LSTM을 SAC와 결합하면 과거의 정보를 반영하여 더 나은 결정을 내릴 수 있음.

모듈을 사용할 경우에 수정해야 하는 값: learning\_starts, batch\_size, learning\_rate, buffer\_size

행동에 대한 보상과 정책 설정에 주의해야 함. (Q값에 민감한 모델이므로 보상 함수 설정 유의)

**실험 설계**

train data: 5-6월 (4,5월과 6, 7월로 구분할 경우에, 최소, 최대 범위가 과하게 다른 문제 발생)

test data: 4월, 7월 (계절성이 여름만 학습되었으므로 일반화 검증 목적)

상태 공간: -np.inf에서 np.inf 사이, (lookback\*특성 수) 개의 정보 전달, 데이터셋을 평탄화해 전달

행동 공간: -1에서 1 사이, 2개의 행동 예측

init(선언 함수)

* 고정: 원본 데이터셋, 내부 온도 예측(모델, scaler\_X, scaler\_Y), 과거 스텝(lookback 길이)
* 변동: 유동(action에 의해 변경될) 데이터셋, 현재 스텝, 누적 개폐도 변화량

reset(초기화 함수)

* init의 변동에 해당하는 변수들을 최초 선언과 동일한 상태로 초기화
* 유동(action에 의해 변경될) 데이터셋: 원본 데이터셋 복사
* 현재 스텝: 과거 스텝(lookback 길이) 복사
* 누적 개폐도 변화량: 0

step(행동에 따른 다음 상태 공간 및 보상 전달 함수)

* 유동 데이터셋의 현재 데이터셋과 직전 데이터셋을 원래 스케일로 되돌려서 보상 계산 후, 전달할 때 재변형
* action을 원래 스케일로 되돌린 후, 원래 스케일의 유동 데이터셋의 해당 위치에 대체
* 변경된 유동 데이터셋으로 다음 내부 온도 예측
* 다음 유동 데이터셋에 현재의 변경된 정보들을 업데이트 후, 평탄화해 다음으로 전달
* 보상

1. 내부 온도 변화량 최소화: 예측 내부 온도와 현재 내부 온도의 차이의 제곱한 만큼 감소
2. 내부 온도 범위 제한: 예측 내부 온도가 15 이하이거나 35 이상일 경우에 1번을 5배로 수정
3. 개폐도 변화량 최소화: action에 의한 개폐도와 이전 개폐도의 차이를 각각 제곱하여 더한 값의 0.01배만큼 감소
4. 장기 개폐도 변화량 최소화: 3번의 값들을 누적해서 추가한 누적 개폐도 변화량의 0.3배만큼 감소
5. 풍향, 풍속에 의한 풍상측 천창 개폐도 제한:
   1. 풍향이 80-100 범위이면서 풍속이 5 이상이고 개폐도가 70 이상이면, action에 의한 풍상측 천창 개폐도와 이전 풍상측 천창 개폐도의 차이의 제곱을 2배로 수정
   2. 풍향에 관계없이 풍속이 6 이상이고 개폐도가 40 이상이면, action에 의한 풍상측 천창 개폐도와 이전 풍상측 천창 개폐도의 차이의 제곱을 2배로 수정

* done: 현재 스텝이 마지막 데이터셋(원본 데이터셋 길이-1)에 해당하면 종료
* info(전달하는 정보): 예측 내부 온도, 현재 스텝(예측 내부 온도의 스텝)

시뮬레이션 기간은 1주일(2016 스텝)로 설정

평가 기준: 실제와 예측 결과 비교

* 그래프

1. 내부 온도 변화 비교
2. 풍하측 천창 개폐도 변화 비교
3. 풍상측 천창 개폐도 변화 비교
4. 내부 온도 변화량 비교
5. 풍하측 천창 개폐도 변화량 비교
6. 풍상측 천창 개폐도 변화량 비교

* 변화량의 합

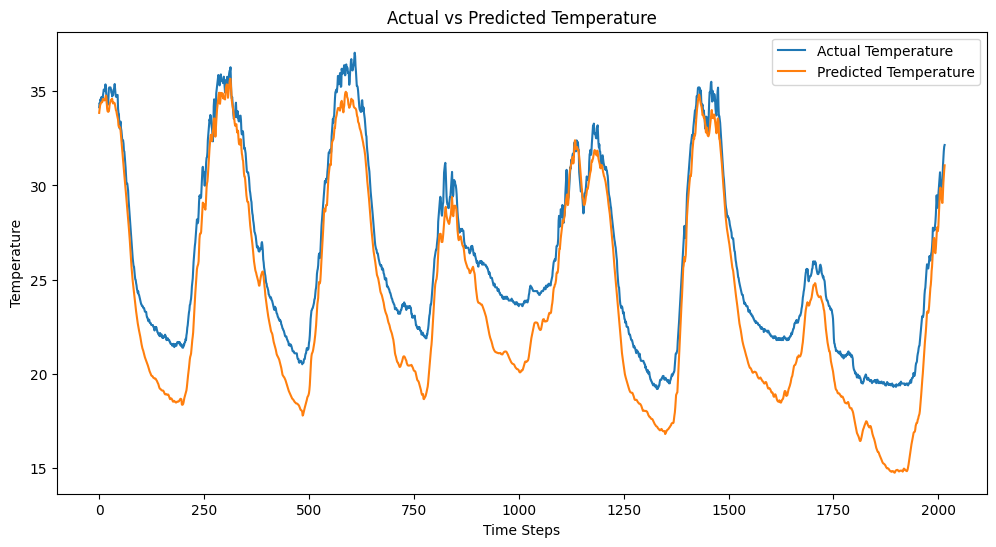
1. 내부 온도 변화량의 합
2. 풍하측 천창 개폐도 변화량의 합
3. 풍상측 천창 개폐도 변화량의 합
4. 전체 천창 개폐도의 합

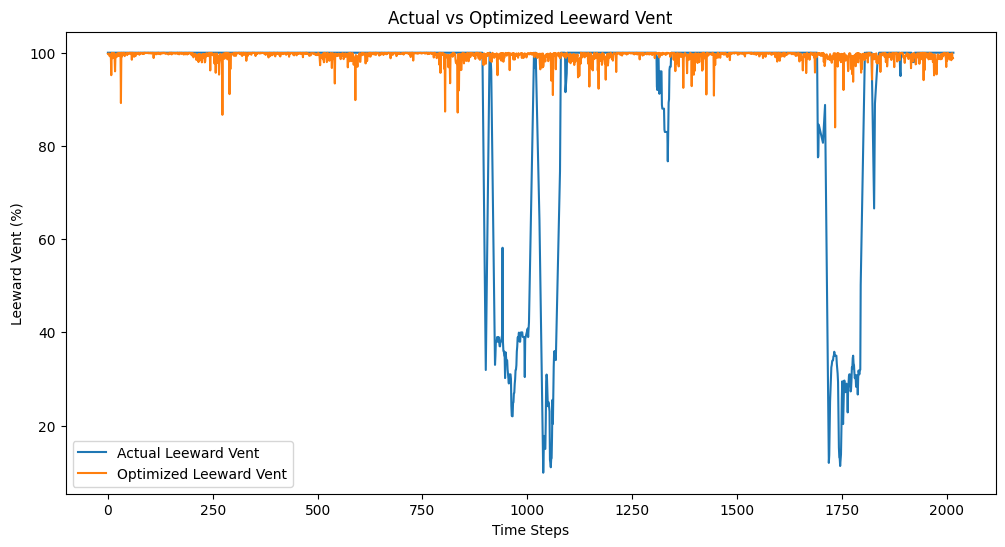
* 그 외 평가 기준

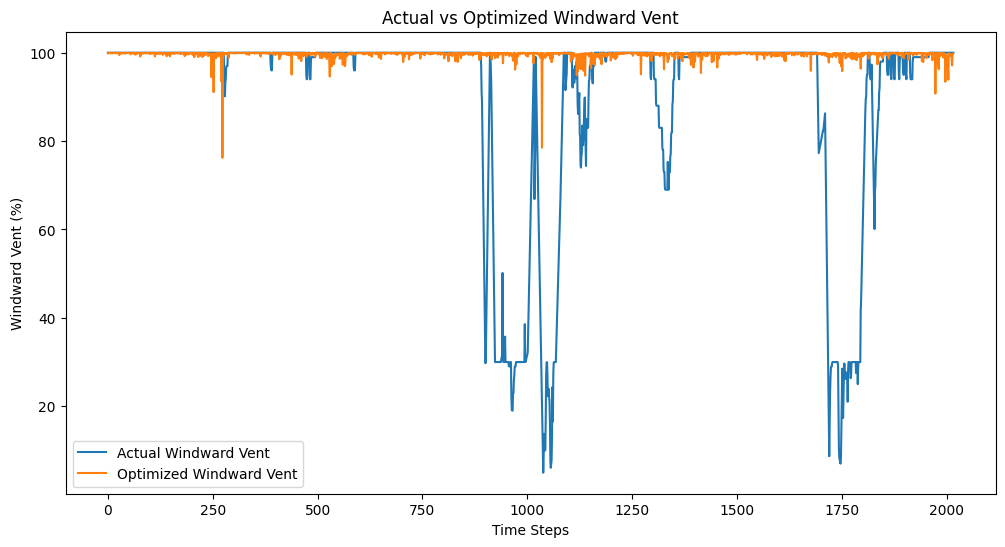
1. 목표 온도 범위(10에서 35 사이) 준수 비율
2. 최대 온도 변화량(0.35) 준수 비율
3. 12시점마다 최대 온도와 최소 온도의 차이 제한(4 이하) 준수 비율

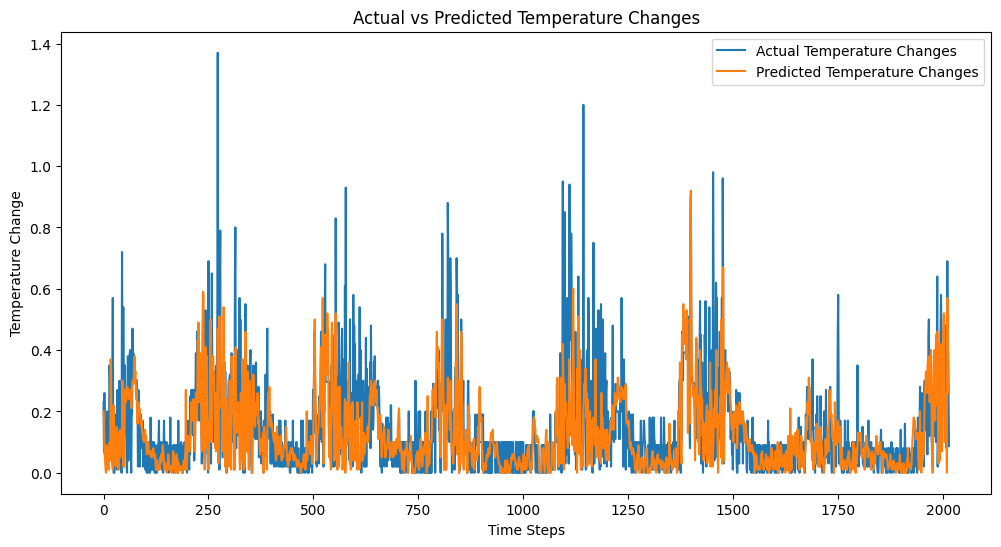
**결과 1**train data(4월)

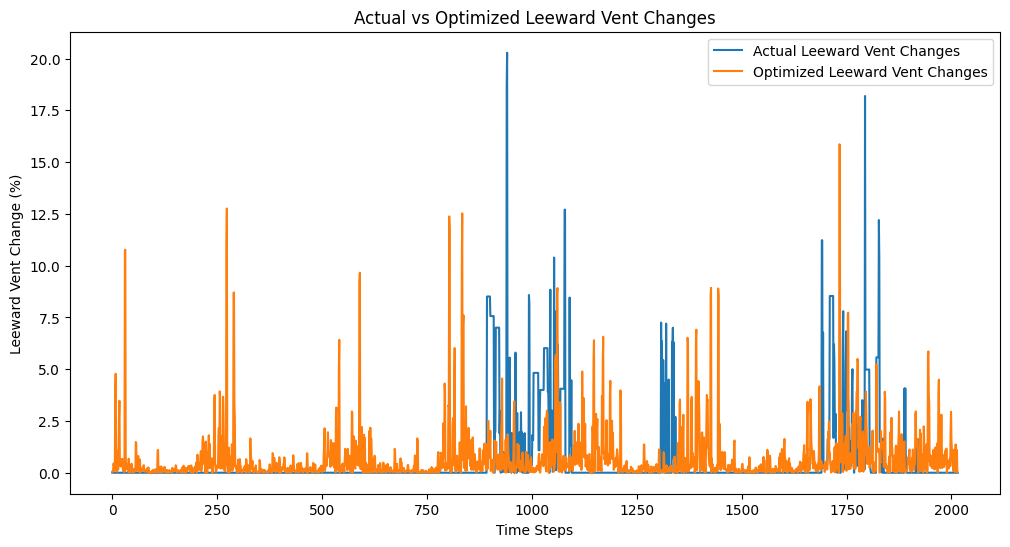
test data(7월)

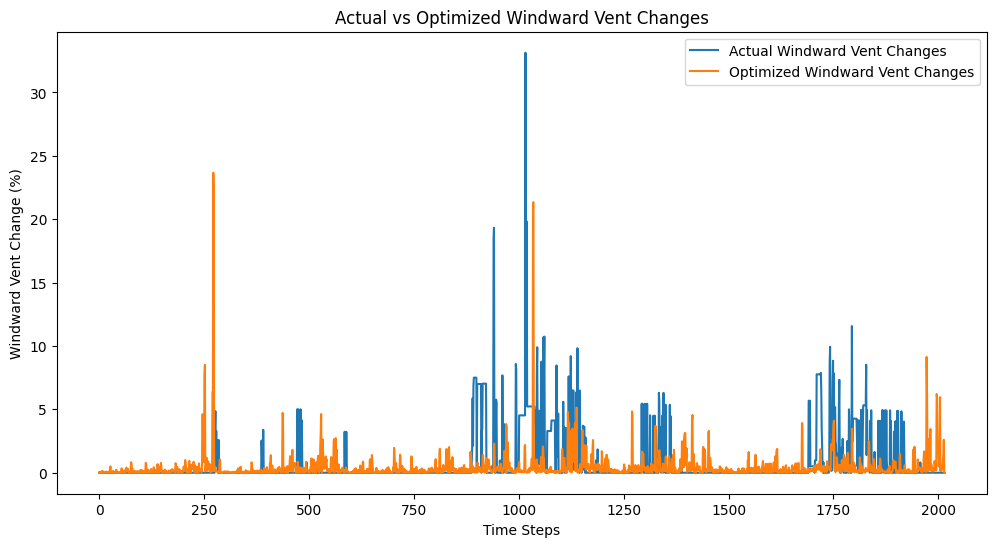












Temperature Change MAE: 0.08744027258850337

Temperature Change MSE: 0.01781714658005741

Temperature Change RMSE: 0.13348088469911118

Leeward Vent Change MAE: 1.0558873423990482

Leeward Vent Change MSE: 5.14581861687535

Leeward Vent Change RMSE: 2.268439687731492

Windward Vent Change MAE: 1.0134606086303946

Windward Vent Change MSE: 5.53357220409207

Windward Vent Change RMSE: 2.352354608491685

Temperature Actual Change Sum: 291.13666666666666

Temperature Predicted Change Sum: 249.22000122070312

Leeward Vent Actual Change Sum: 1131.8999999999999

Leeward Vent Predicted Change Sum: 1436.726448059082

Windward Vent Actual Change Sum: 1499.6999999999998

Windward Vent Predicted Change Sum: 940.7540512084961

Vent Actual Change Sum: 2631.5999999999995

Vent Predicted Change Sum: 2377.480499267578

Temperature Stability Ratio: 0.97

Temperature Change Compliance Ratio: 0.94

Temperature Variation Compliance Ratio: 0.99

Actual Temperature Stability Ratio: 0.94

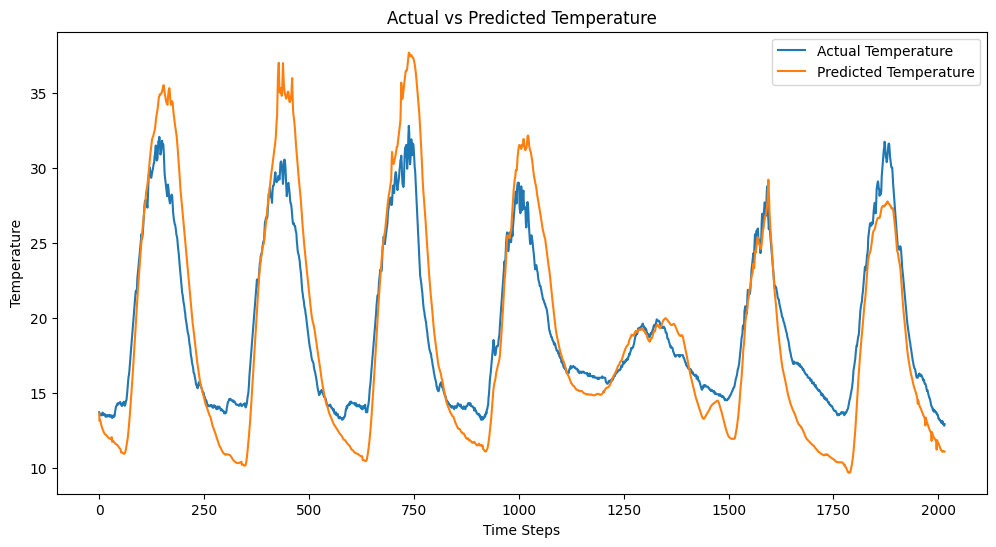
Actual Temperature Change Compliance Ratio: 0.90

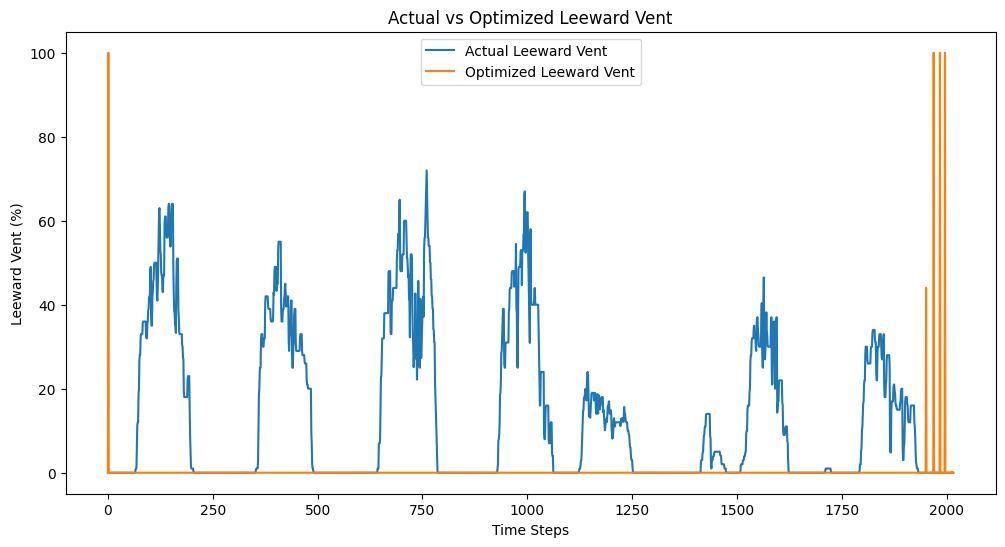
Actual Temperature Variation Compliance Ratio: 0.99

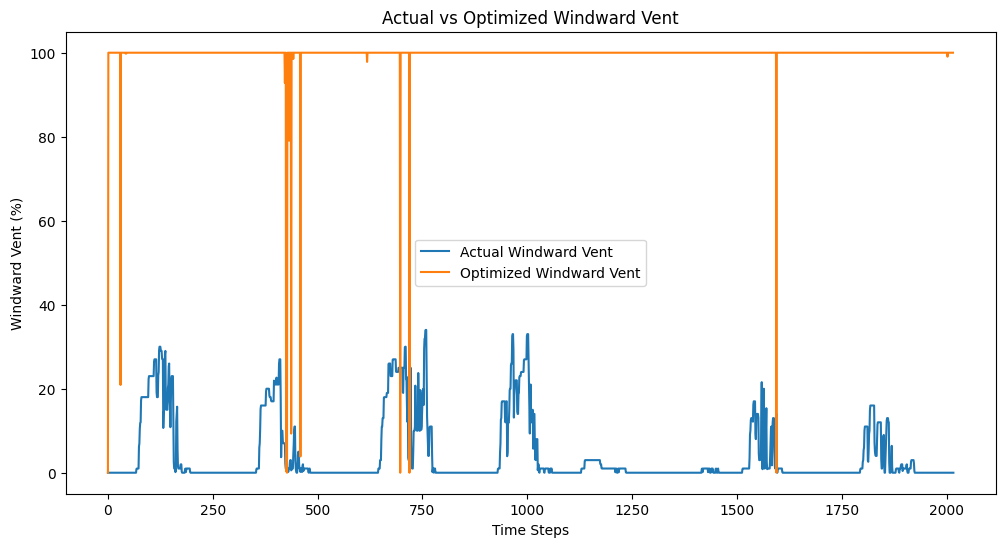
**결과 2**

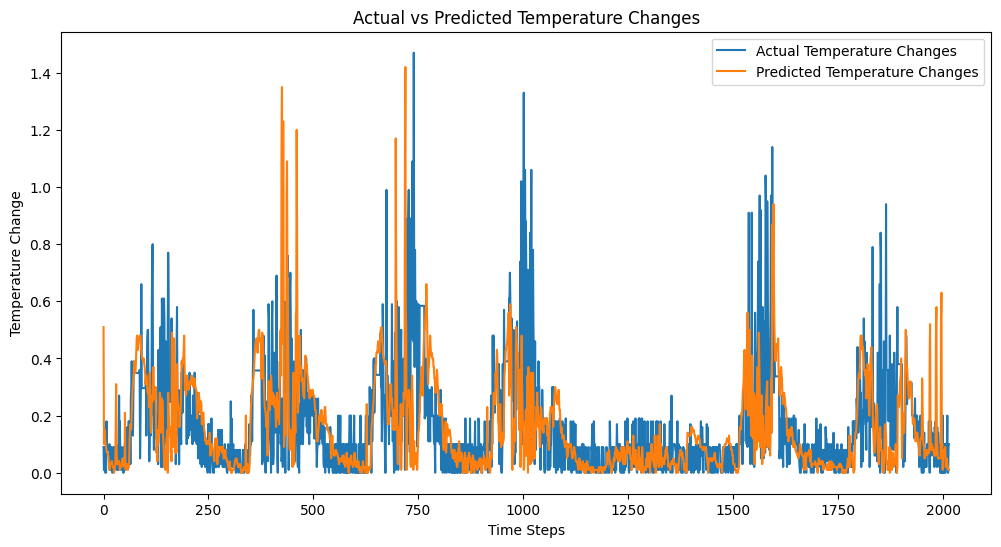
train data(5, 6월)

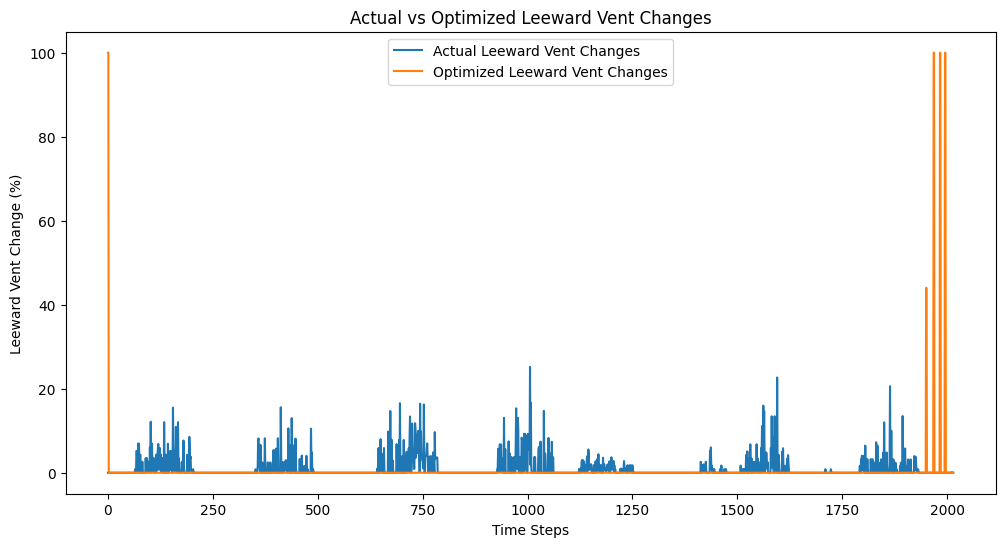
test data(4월)

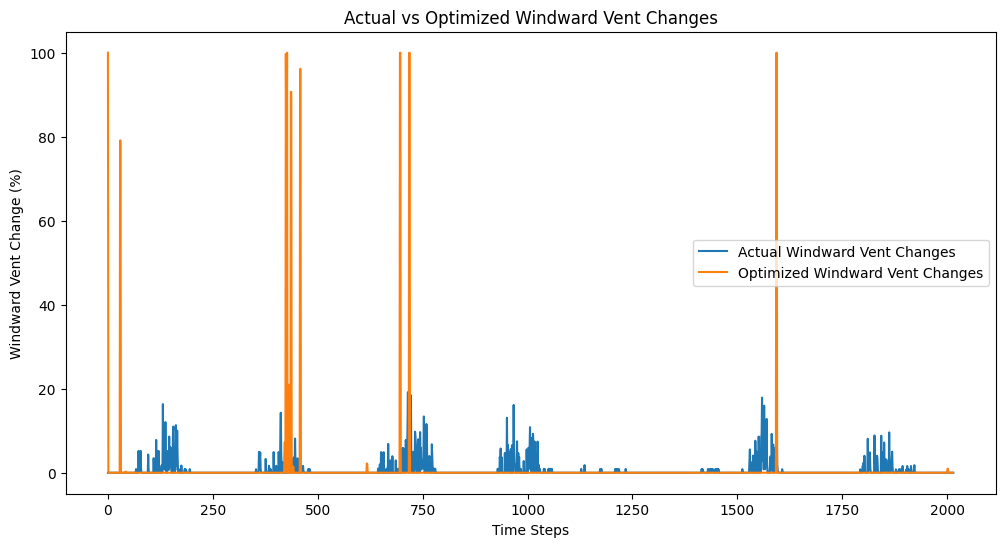












Temperature Actual Change Sum: 334.65999999999997

Temperature Predicted Change Sum: 316.0

Leeward Vent Actual Change Sum: 1994.4600000000003

Leeward Vent Predicted Change Sum: 888.419068813324

Windward Vent Actual Change Sum: 1240.8600000000001

Windward Vent Predicted Change Sum: 1653.5985808372498

Vent Actual Change Sum: 3235.3200000000006

Vent Predicted Change Sum: 2542.0176496505737

Temperature Stability Ratio: 0.96

Temperature Change Compliance Ratio: 0.88

Temperature Variation Compliance Ratio: 0.91

Actual Temperature Stability Ratio: 1.00

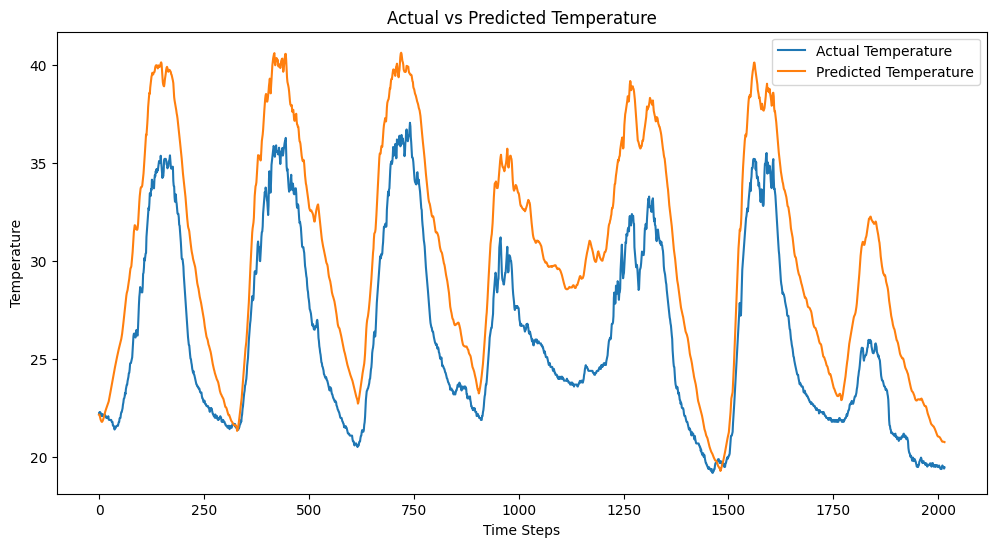
Actual Temperature Change Compliance Ratio: 0.87

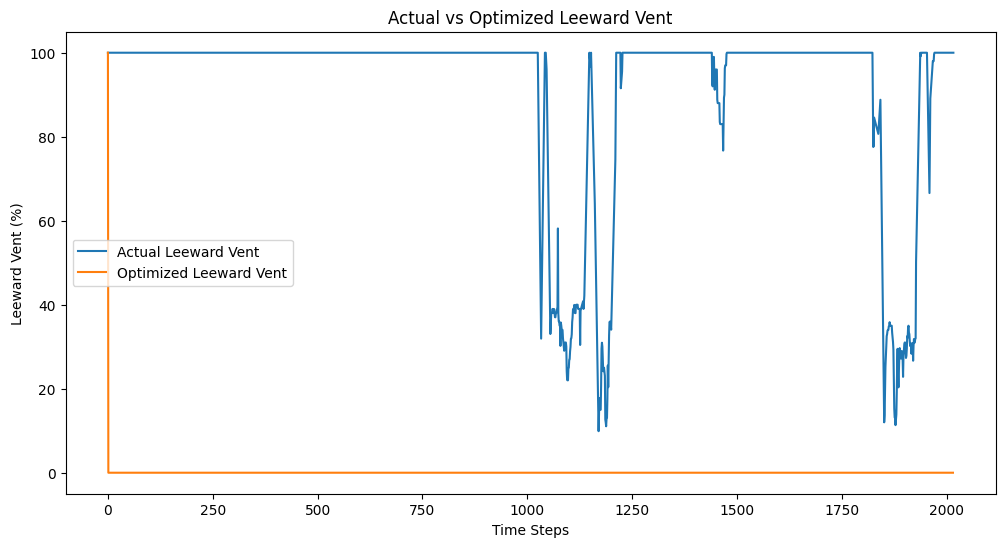
Actual Temperature Variation Compliance Ratio: 0.99

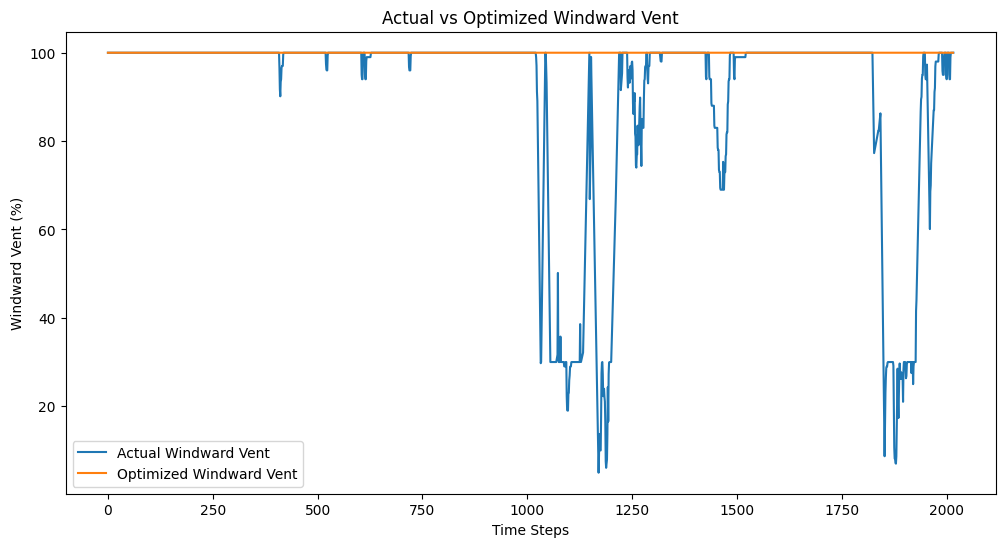
**결과 3**

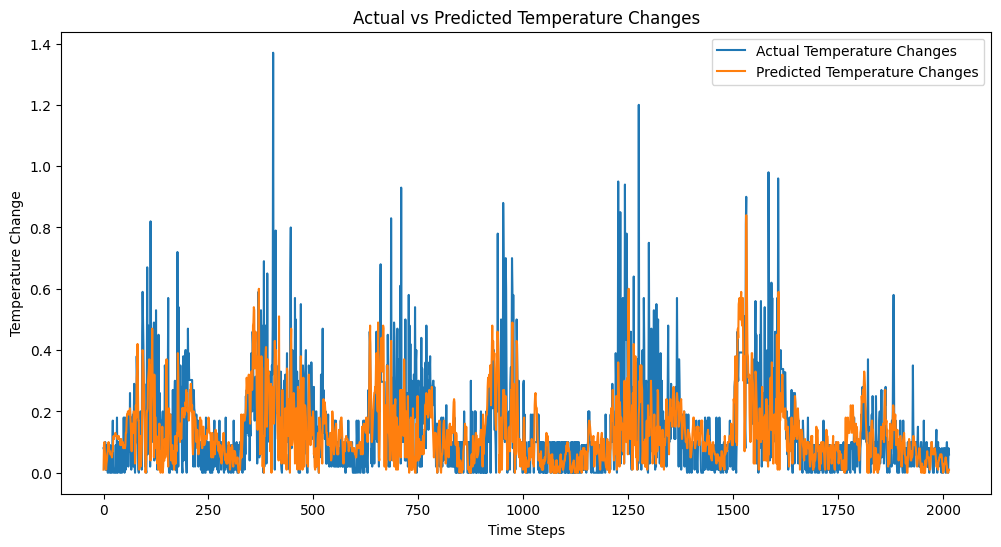
train data(5, 6월)

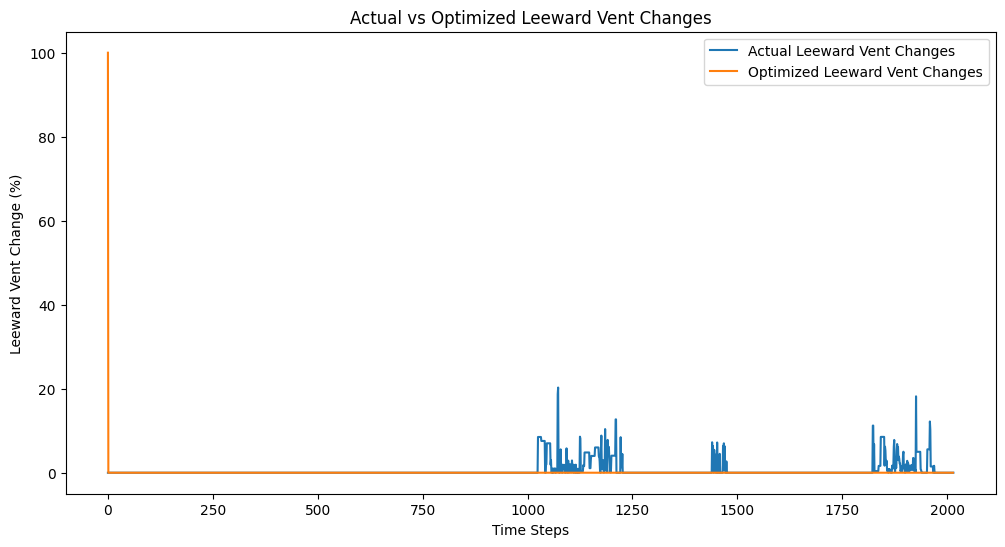
test data(7월)

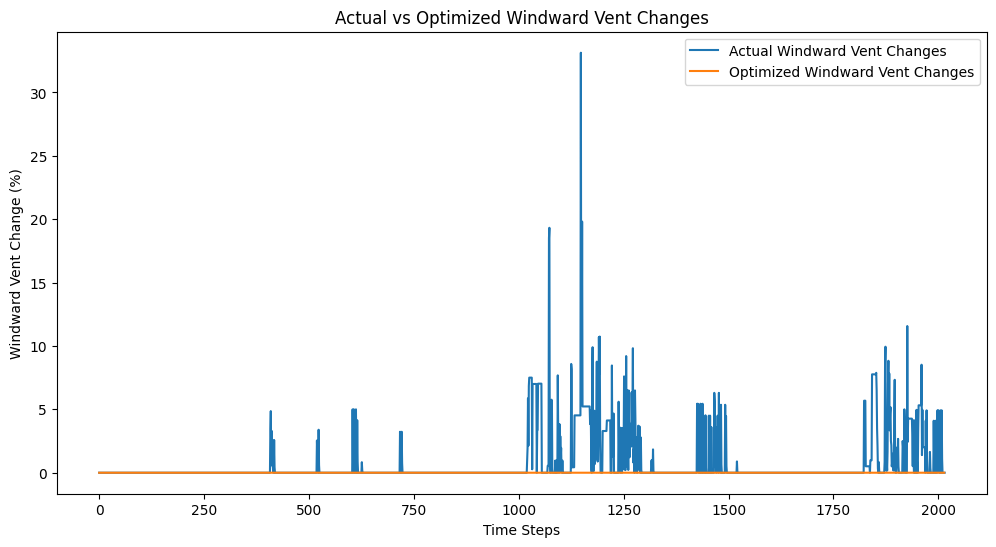












Temperature Actual Change Sum: 289.09000000000003

Temperature Predicted Change Sum: 261.47998046875

Leeward Vent Actual Change Sum: 1117.9

Leeward Vent Predicted Change Sum: 100.0

Windward Vent Actual Change Sum: 1455.7

Windward Vent Predicted Change Sum: 0.0

Vent Actual Change Sum: 2573.6000000000004

Vent Predicted Change Sum: 100.0

Temperature Stability Ratio: 0.74

Temperature Change Compliance Ratio: 0.95

Temperature Variation Compliance Ratio: 0.98

Actual Temperature Stability Ratio: 0.94

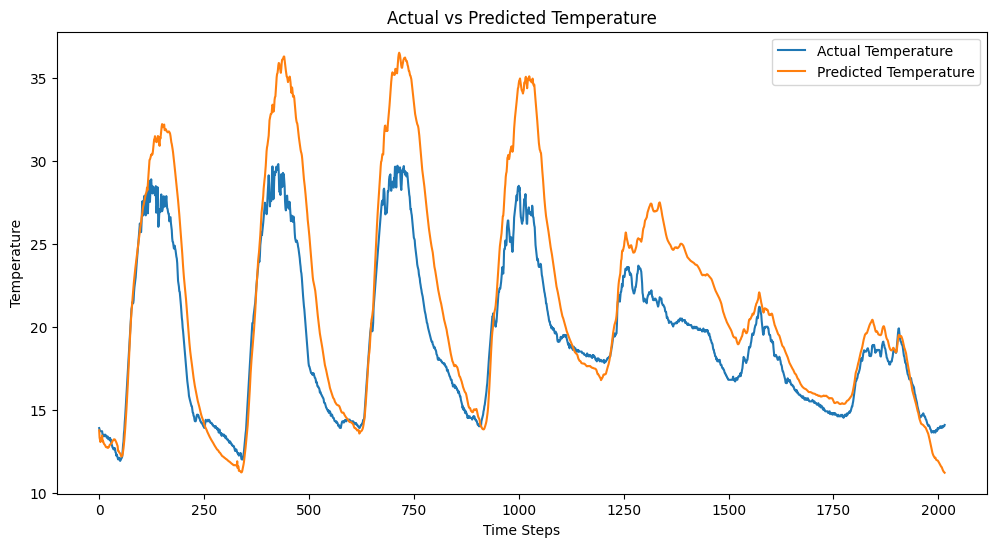
Actual Temperature Change Compliance Ratio: 0.91

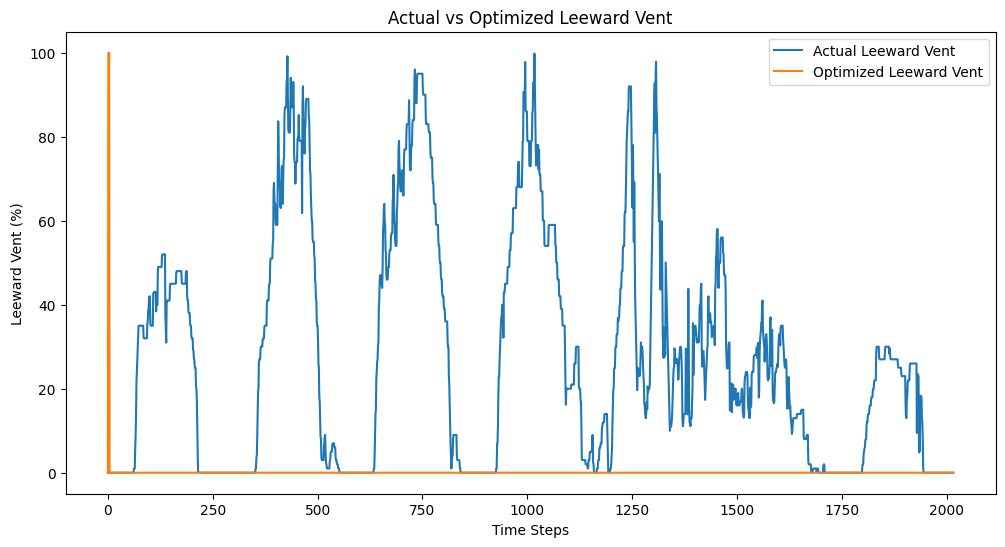
Actual Temperature Variation Compliance Ratio: 0.99

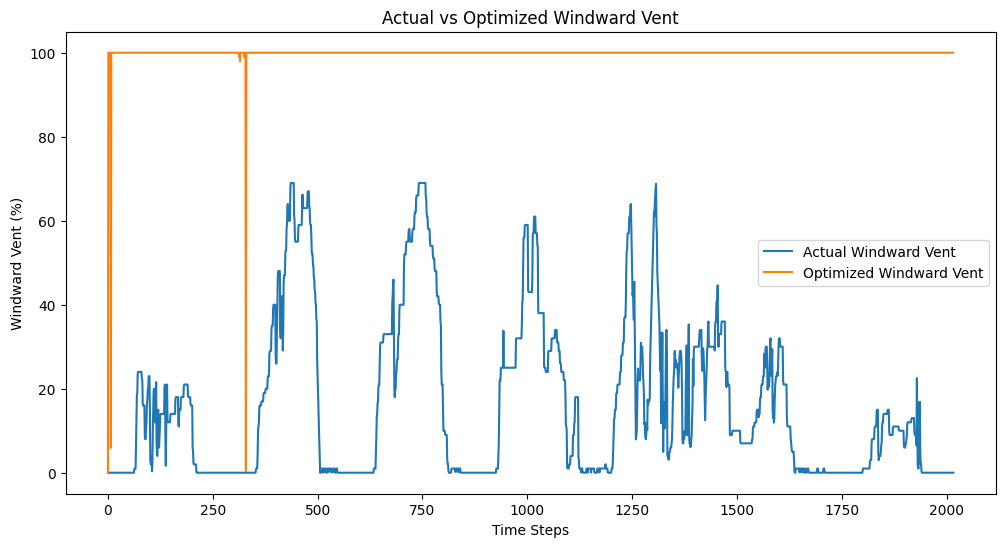
**결과 5-과적합 확인 목적**

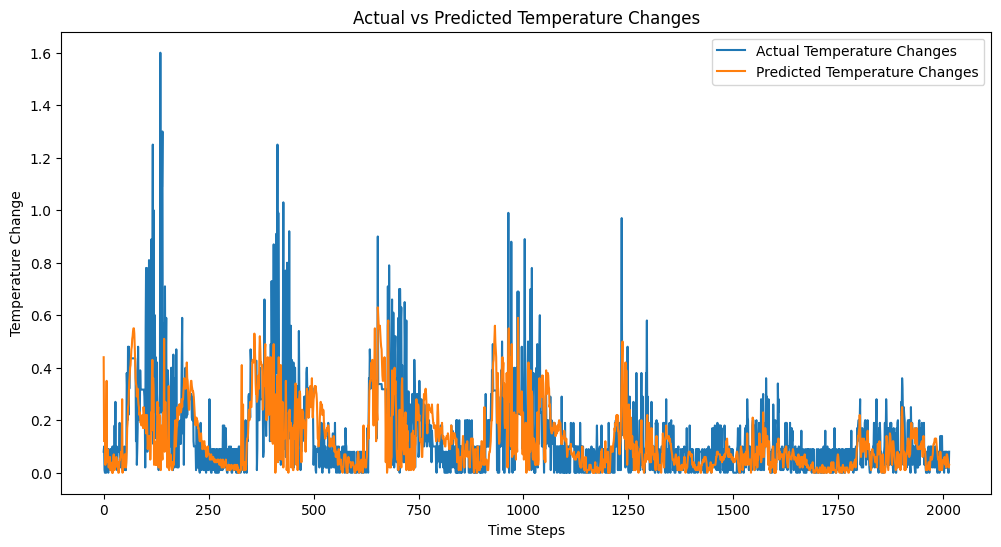
train data(5, 6월)

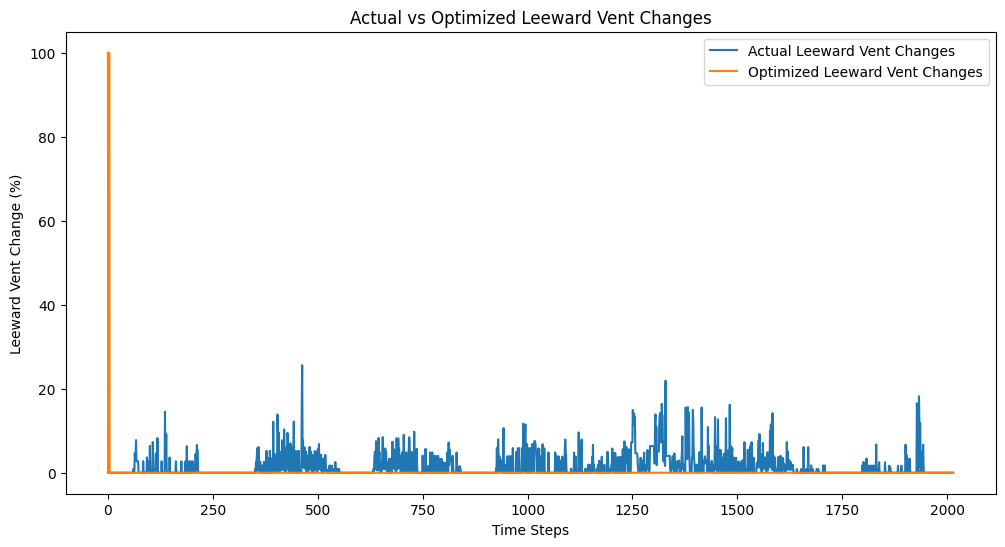
test data(5월)

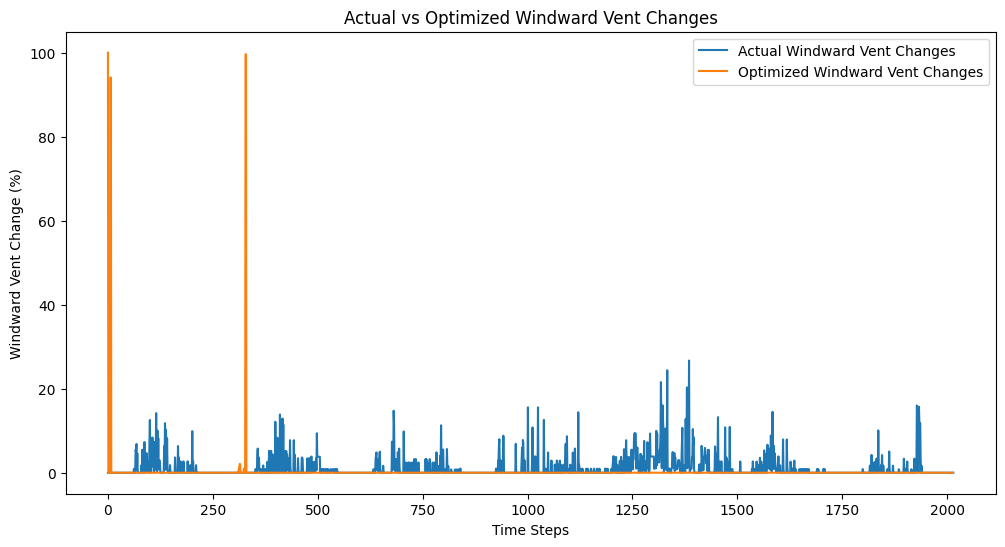












Temperature Actual Change Sum: 288.66999999999996

Temperature Predicted Change Sum: 248.23001098632812

Leeward Vent Actual Change Sum: 2909.76

Leeward Vent Predicted Change Sum: 200.0

Windward Vent Actual Change Sum: 2132.42

Windward Vent Predicted Change Sum: 495.49743127822876

Vent Actual Change Sum: 5042.18

Vent Predicted Change Sum: 695.4974312782288

Temperature Stability Ratio: 0.96

Temperature Change Compliance Ratio: 0.93

Temperature Variation Compliance Ratio: 0.95

Actual Temperature Stability Ratio: 1.00

Actual Temperature Change Compliance Ratio: 0.91

Actual Temperature Variation Compliance Ratio: 0.99