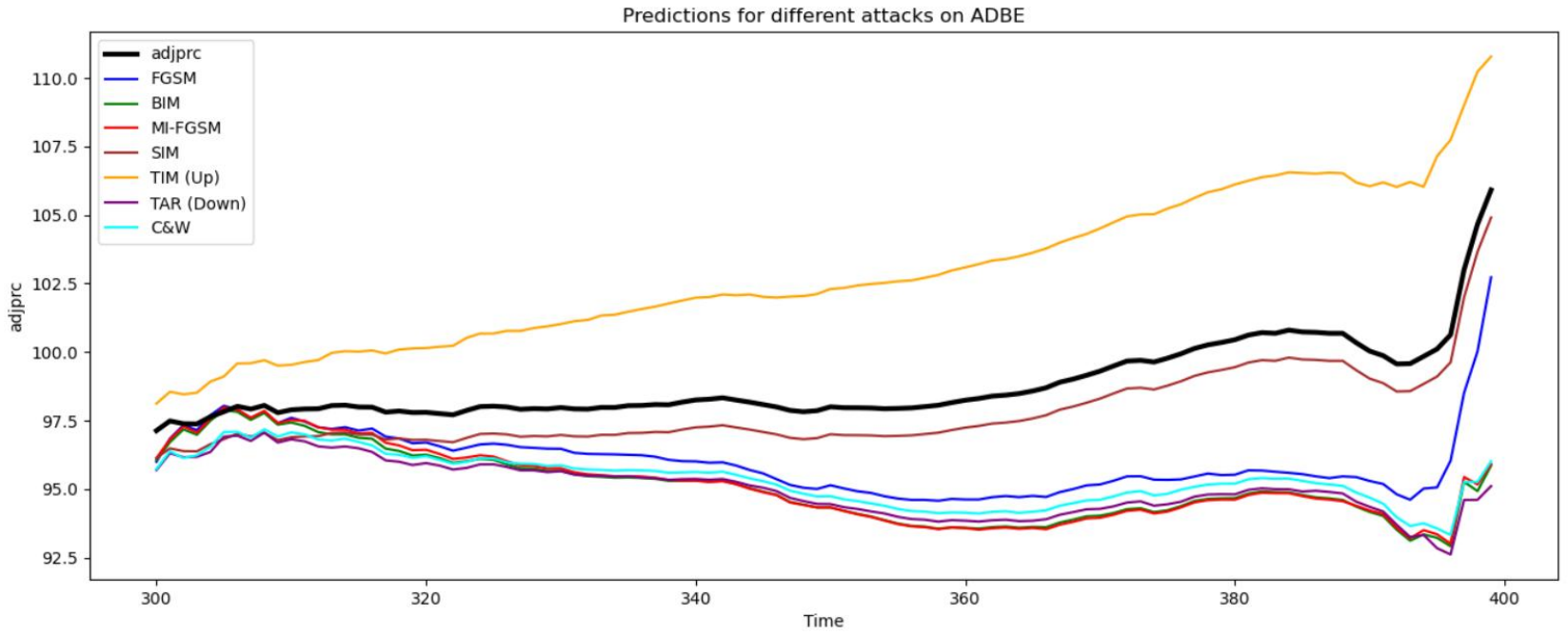
CSCD94 Week 10 Update

Dominik Luszczynski

Slope-Based Attacks

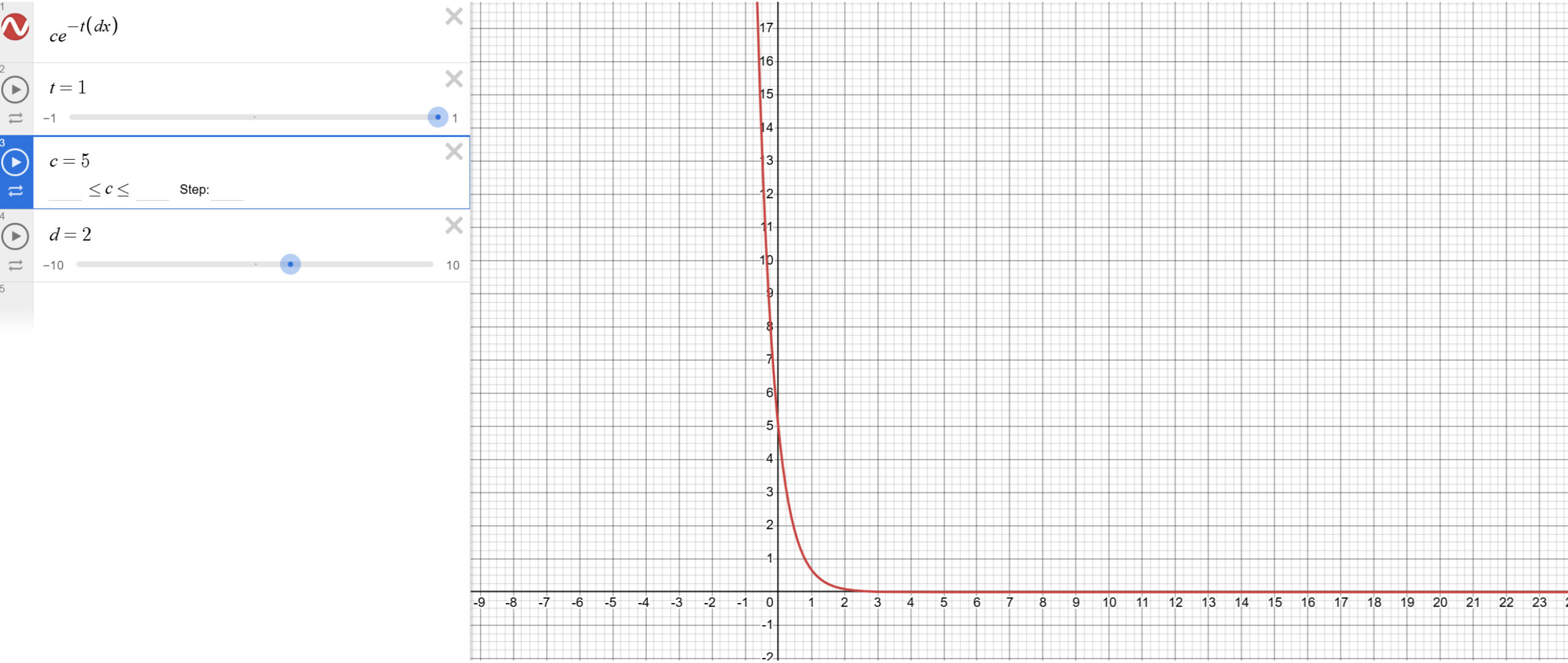
- As we have seen in previous weeks, current versions of targeted attacks tend to simply move the entire forecast above or below the true adjusted price.



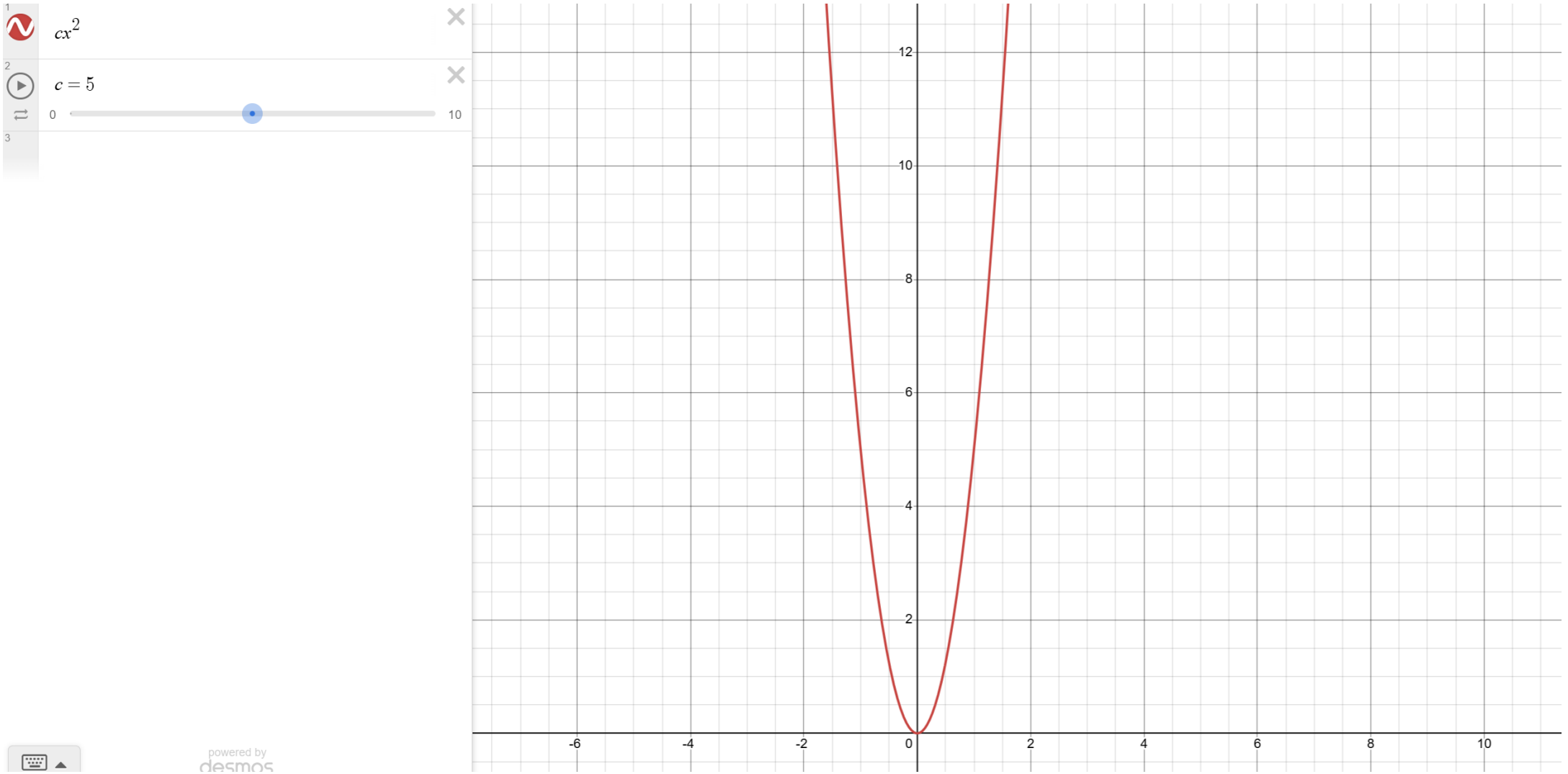
Slope-Based Attacks

- Rather than choosing a target adjprc, and reducing the loss between the target and the forecasted predictions, we instead:
  - Define a slope-based loss function based on a target direction
  - Try to minimize this loss function by creating x_adv (same algorithm as TIM/BIM)
  - Clip x_adv to be within the epsilon-neighbourhood

Sloped-Based Loss Function



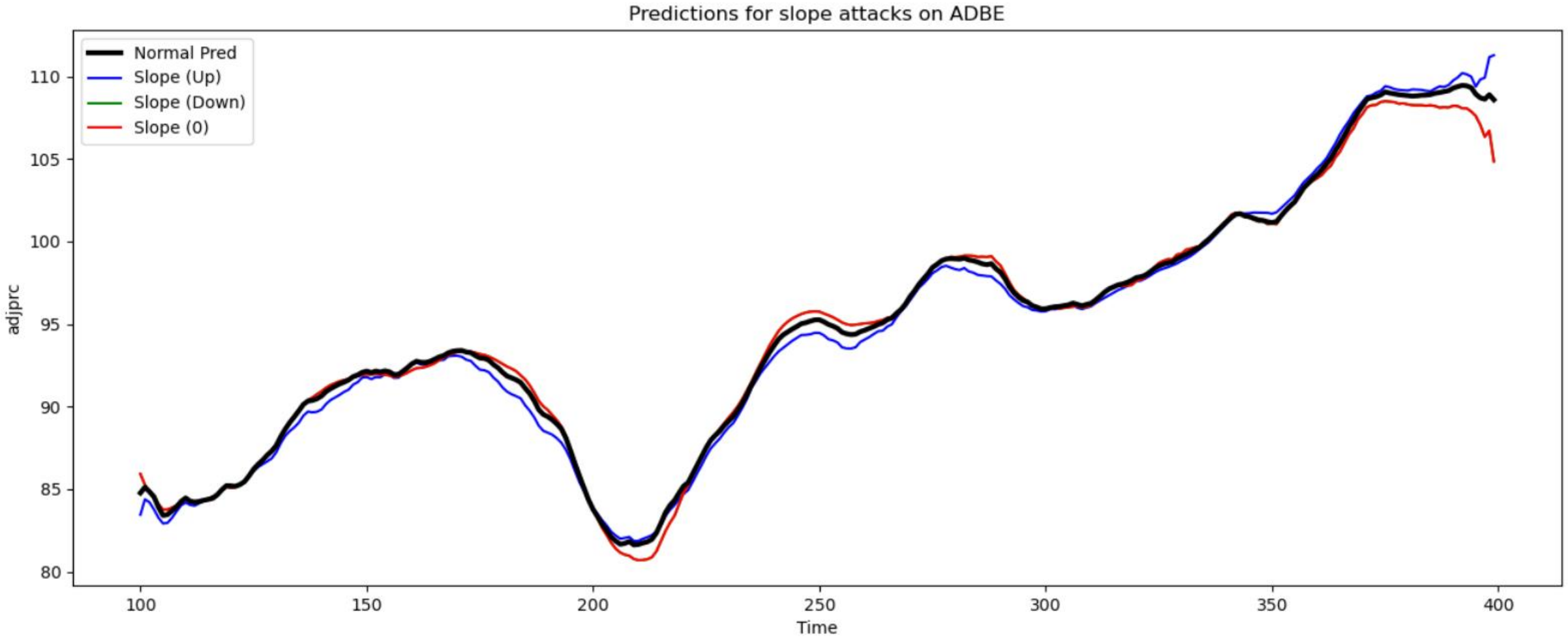
Slope-based Loss Function (slope = 0)



Algorithm

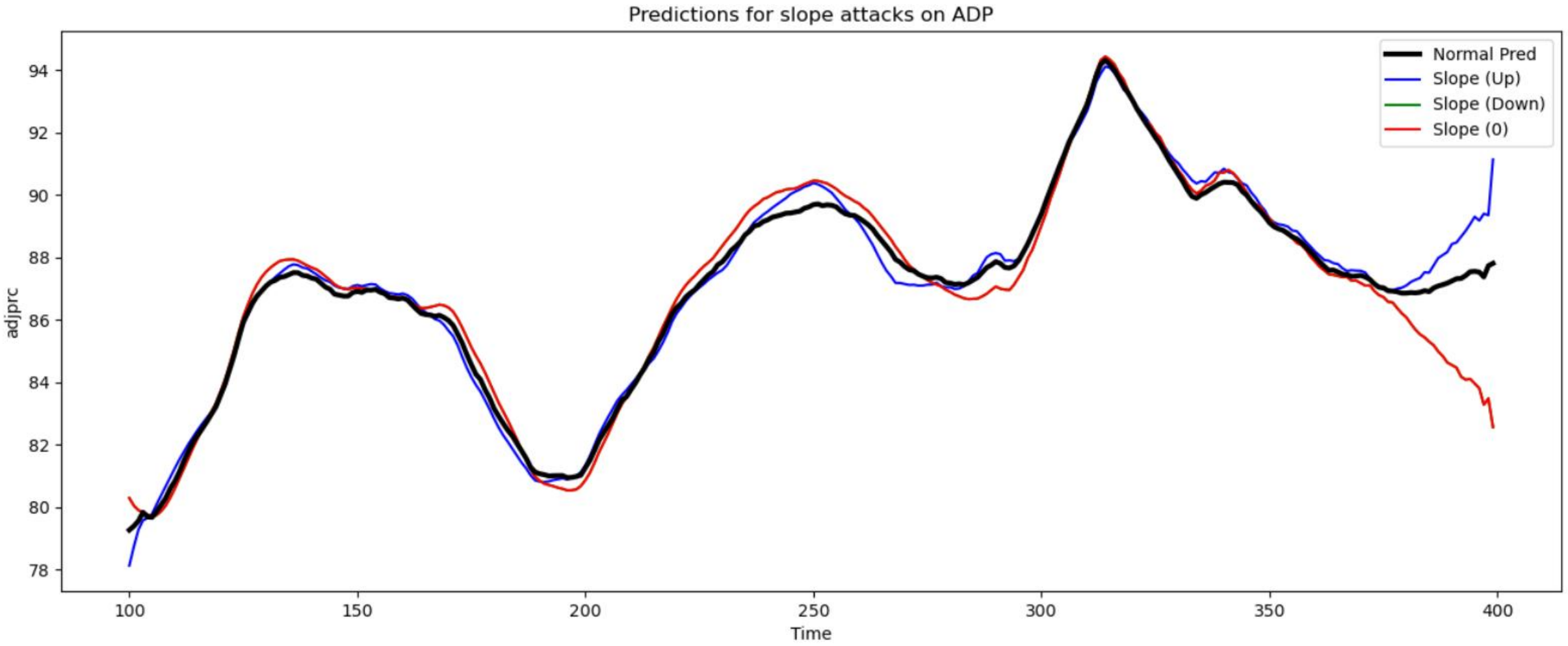
1. x_i = adjprc
2. For i in range of num_iterations:
    1. x_i.requires_grad_(True)
    2. model.zero_grad()
    3. pred = get_predictions(x_i)
    4. slope = (pred[-1] – pred[0]) / len(pred)
    5. loss = slope_loss(slope)
    6. loss.backward()
    7. With no gradients:
        1. grad = loss.grad.data
        2. sign_grad = grad.sign()
        3. noise = step_size * sign_grad
        4. x_i = x_i – noise # Want to move in direction of gradient to minimize the loss
        5. x_i = clamp(x_i, adjprc – epsilon, adjprc + epsilon)
    8. x_i.detach()

Slope-Based Attack Results



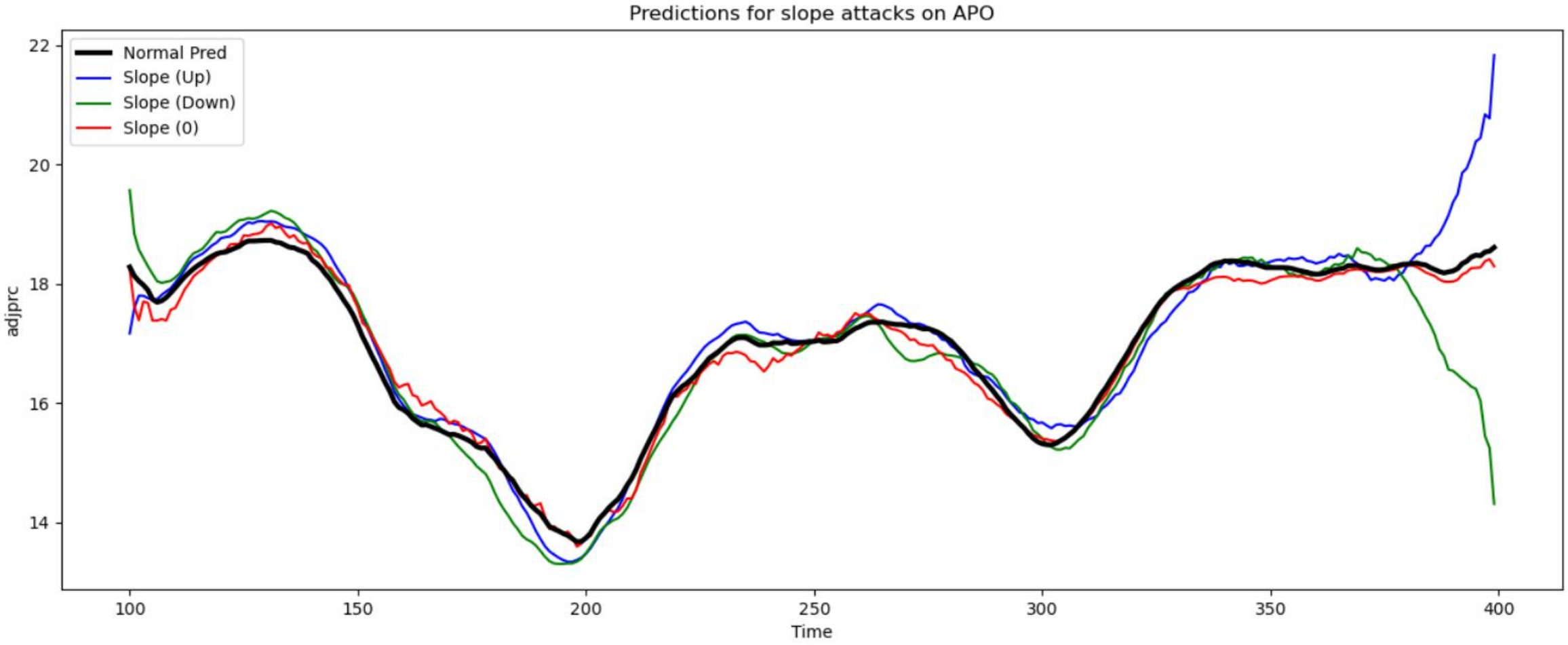
Note: Given that Adjprc tends to increase, Slope (Down) and Slope (0) tend to produce the same output.

Slope-Based Attack Results

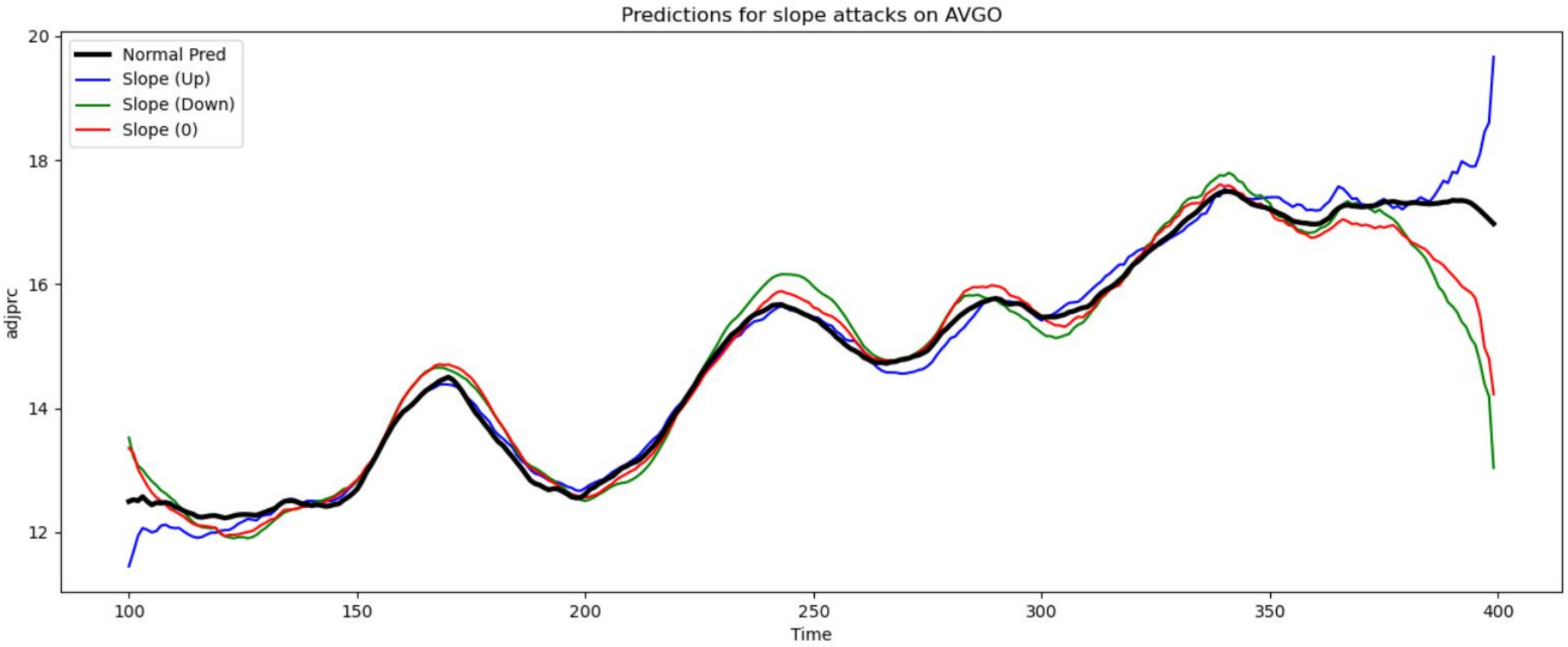


Note: Given that Adjprc tends to increase, Slope (Down) and Slope (0) tend to produce the same output.

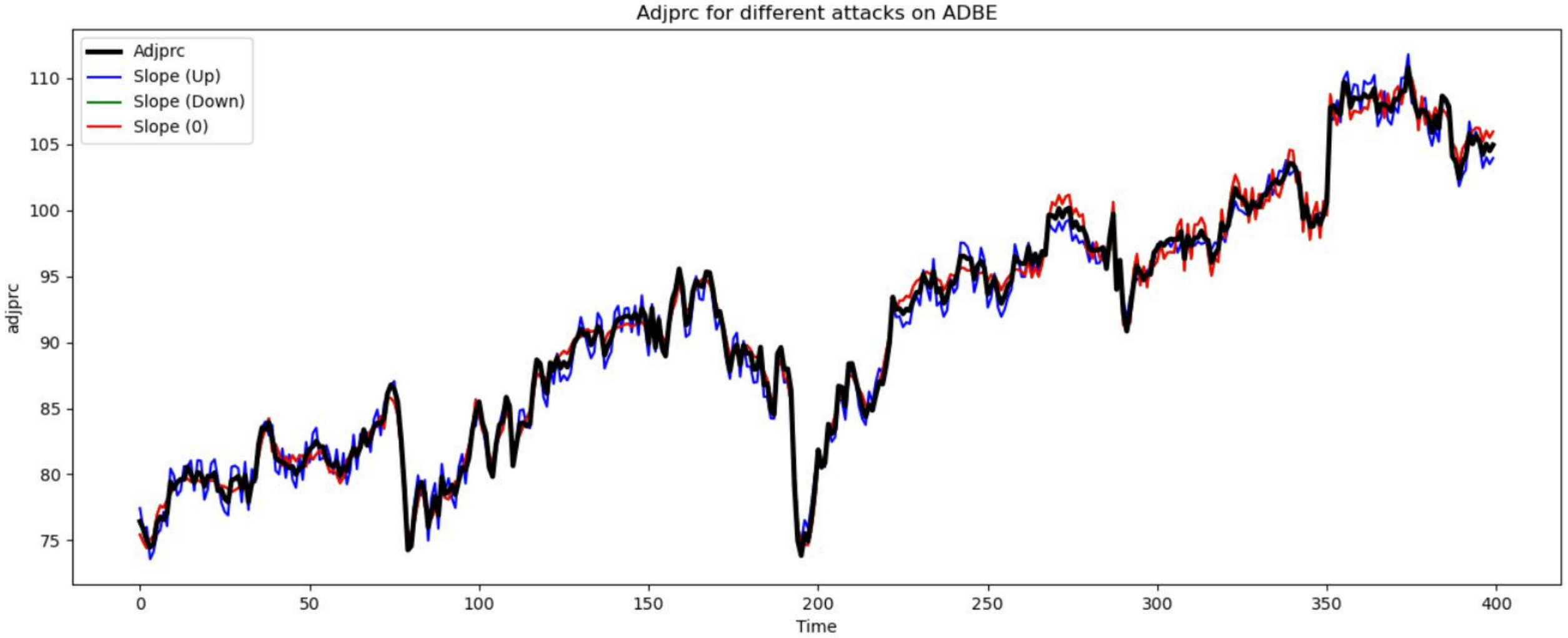
Slope-Based Attack Results



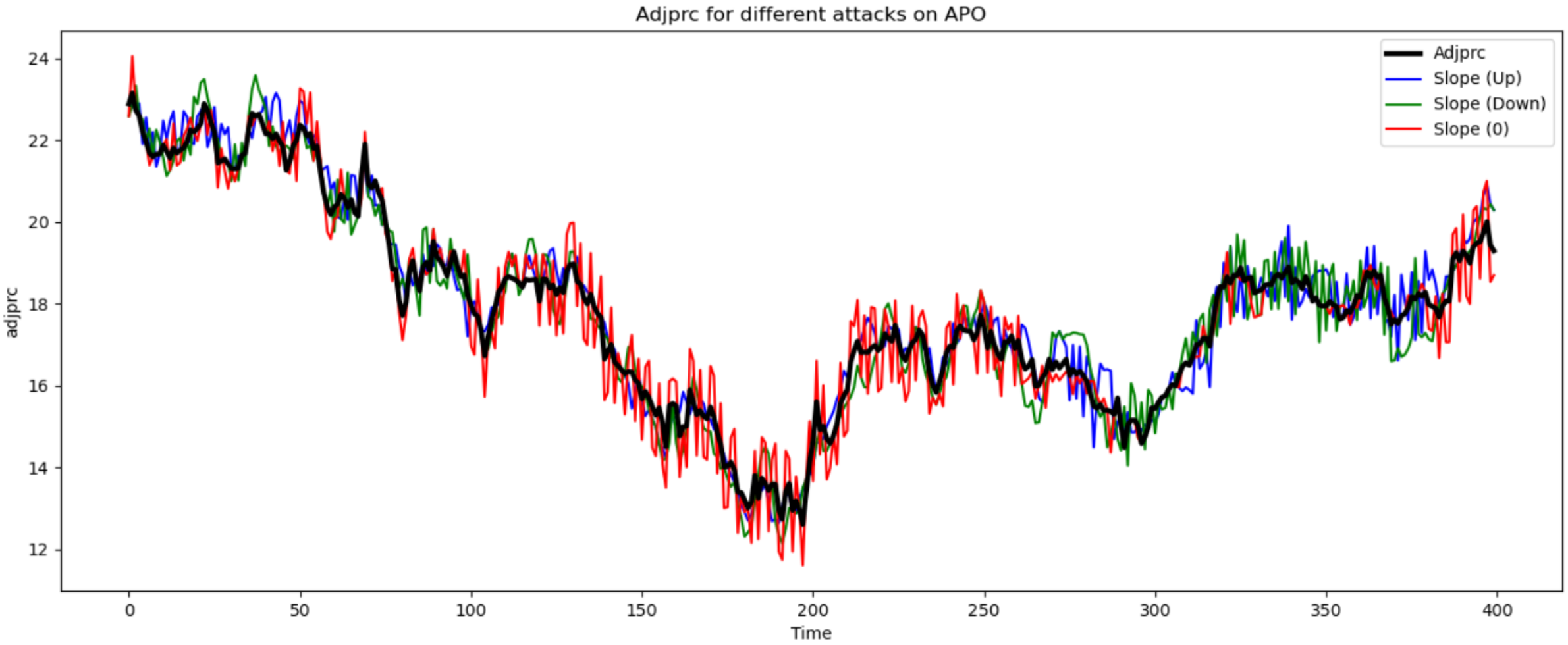
Slope-Based Attack Results



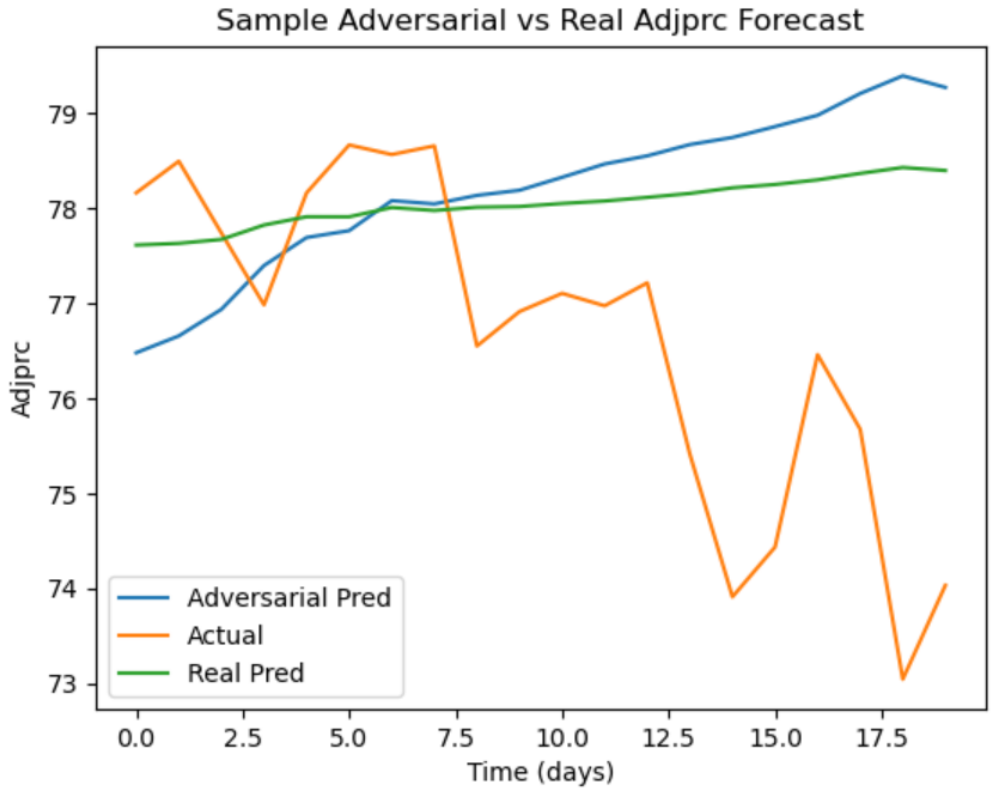
Slope-Based Attack Adjprc

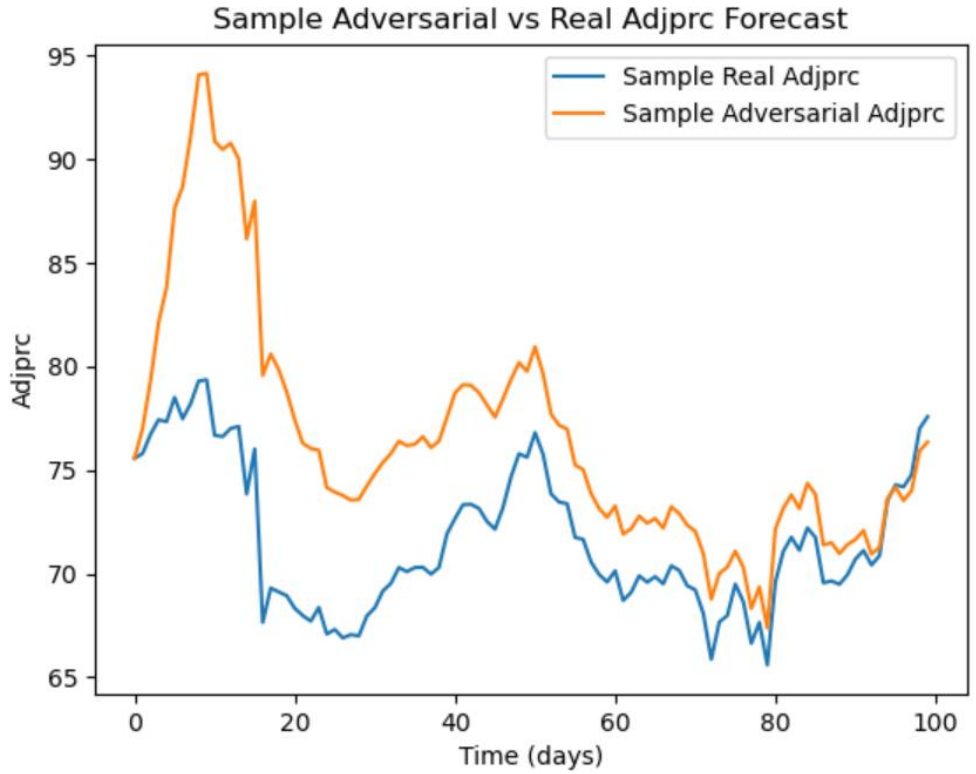


Slope-Based Attack Adjprc

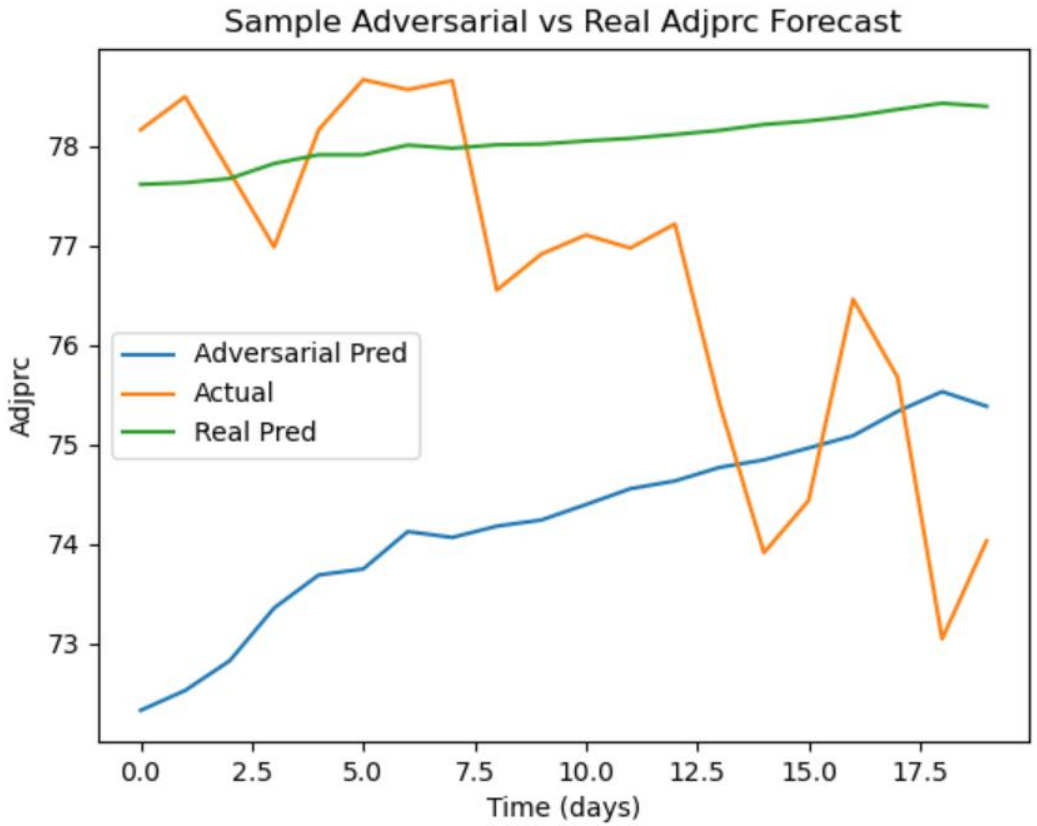


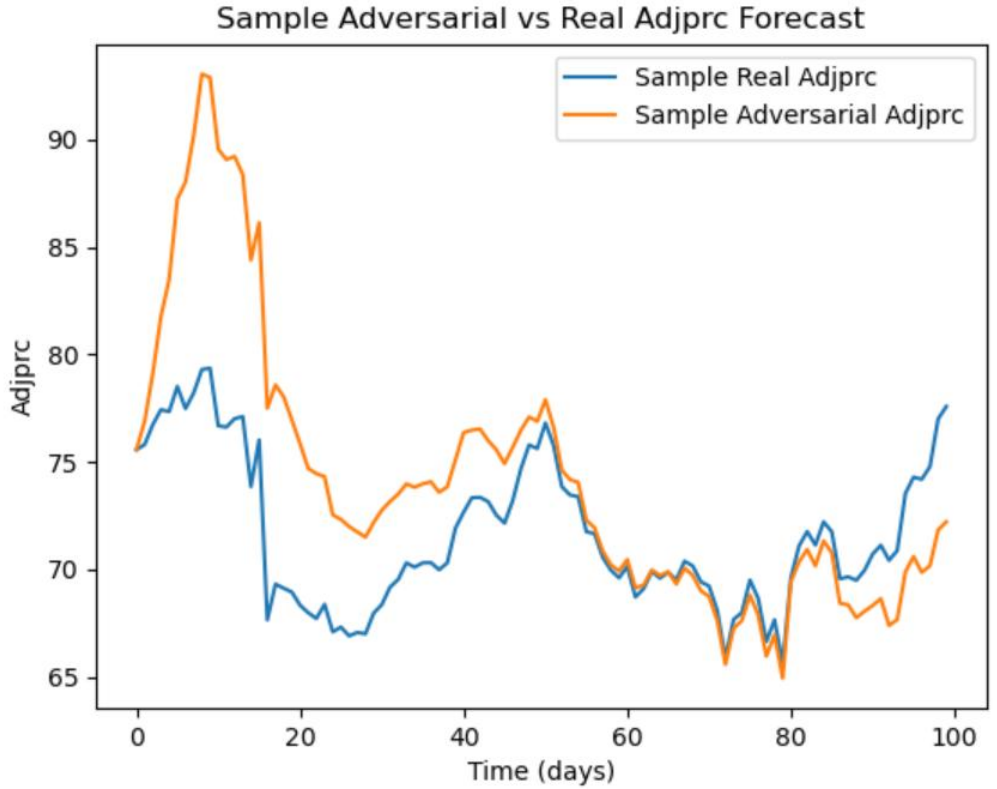
Adapting the Slope-Based Attack into the GAN (Doesn’t work too well)



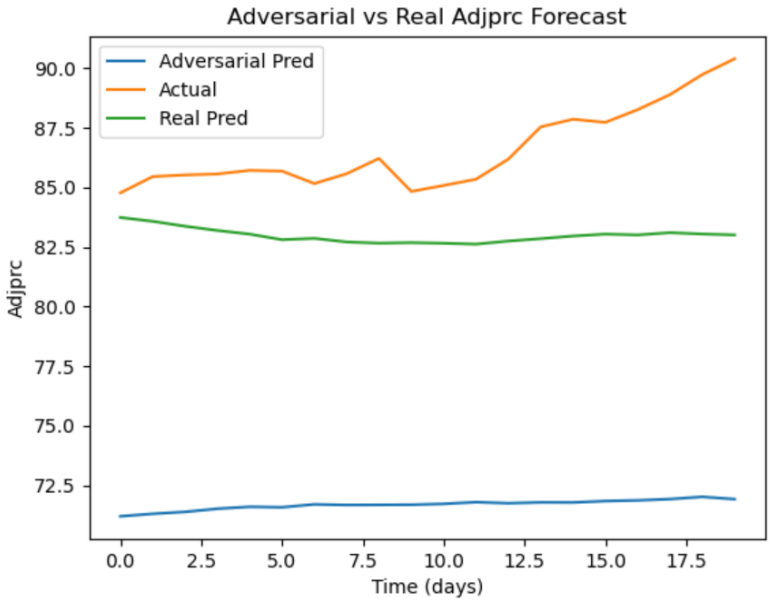


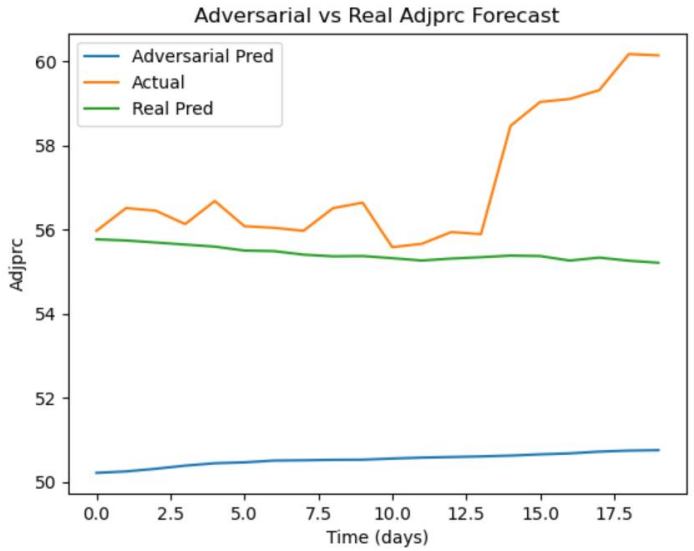
Adapting the Slope-Based Attack into the GAN (Doesn’t work too well)

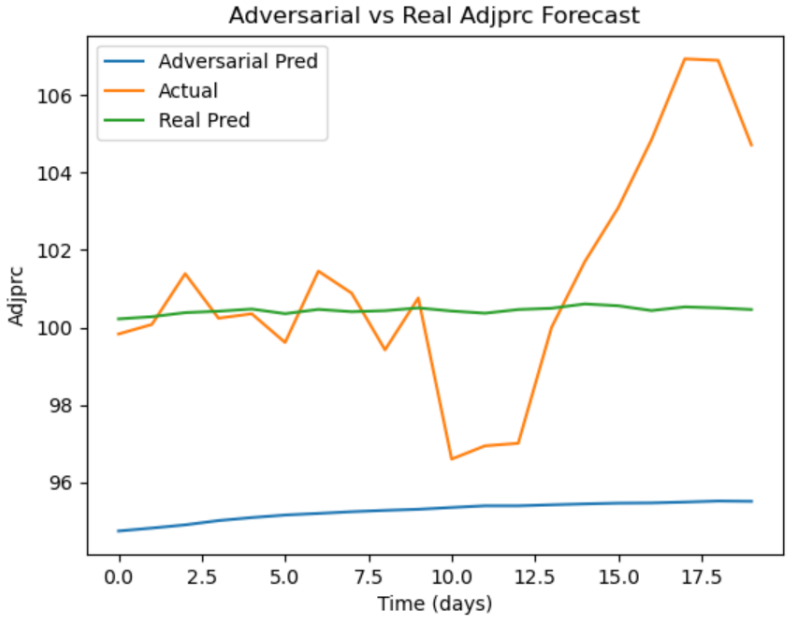




Adapting the Slope-Based Attack into the GAN (Doesn’t work too well)





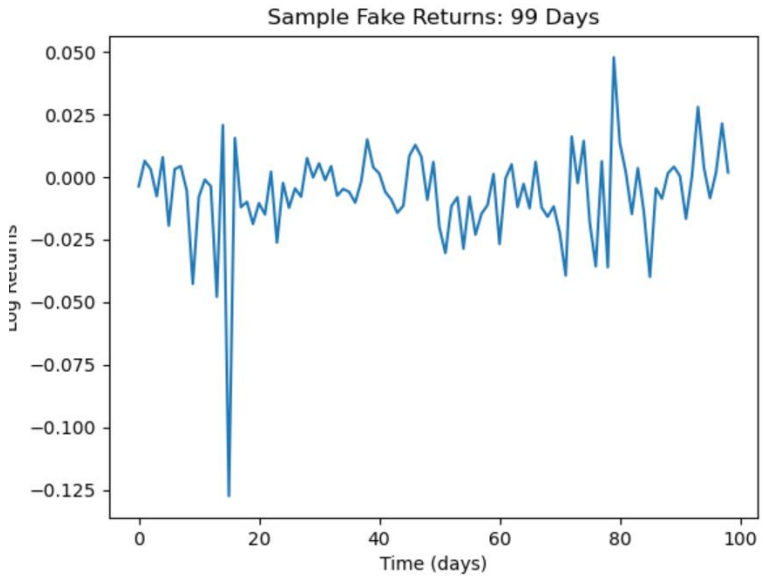


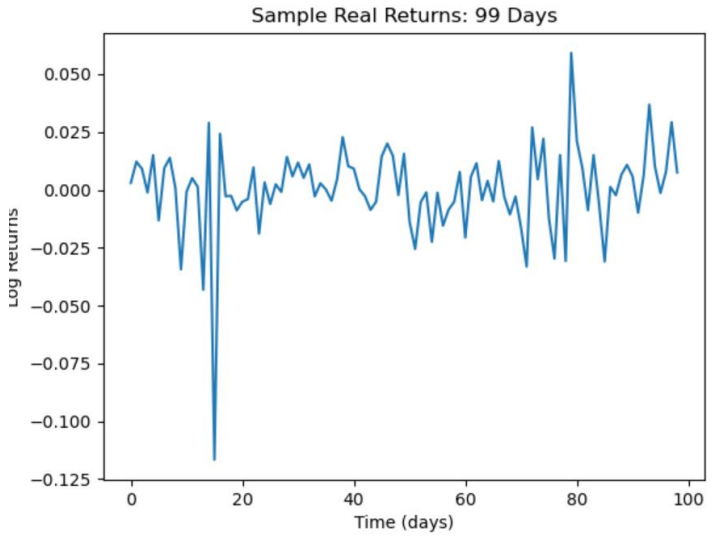
Possible Problems/Fixes

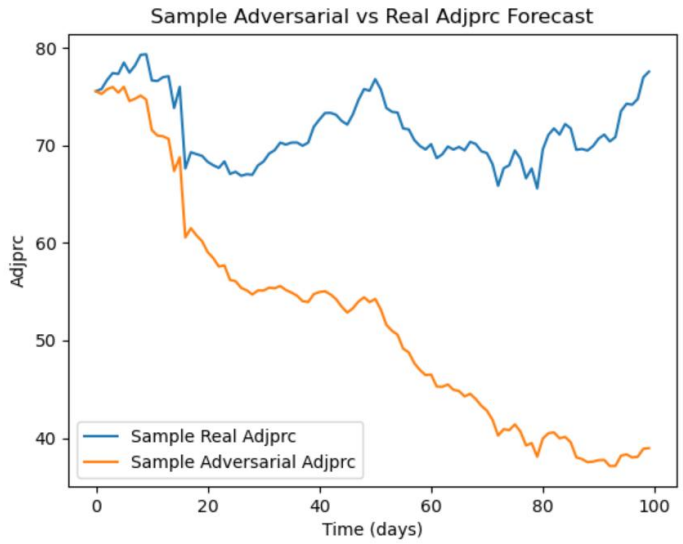
- Currently I have a saved GAN which I preload; however, the critic has probably learnt that the generated sequence should match the condition, thus when the generator tries to make perturbations, the critic “flags” it and gives a high loss to the generator.
- A 20-day forecast may not be enough time to get a meaningful slope, as the Slope-Based Attack generates the attack over a 300-day forecast.

What About Just Maximizing MAE?

- The generator always learns to simply change the mean of the log return







Could add a “Mean-loss” to the generator, like I did for the slope-based version of the GAN

Next Steps

- Experiment with different training set-ups for the GANs
  - Possibly generate longer segments
  - Recursively generate segments to have a large enough forecast for a meaningful slope
  - Add similarity losses such as mean-losses if maximizing MAE.

- Perform the epsilon experiment on the Slope-Based Attacks and generate meaningful comparisons with the TIM/Stealthy Attacks.
  - Also look for ways to make the attacks more stealthy/less noticeable.