Adversarial Observations in Weather Forecasting

Erik Imgrund
BIFOLD & TU Berlin
Berlin, Germany

Thorsten Eisenhofer
BIFOLD & TU Berlin
Berlin, Germany

Konrad Rieck
BIFOLD & TU Berlin
Berlin, Germany

Abstract

AI-based systems, such as Google's GenCast, have recently redefined the state of the art in weather forecasting, offering more accurate and timely predictions of both everyday weather and extreme events. While these systems are on the verge of replacing traditional meteorological methods, they also introduce new vulnerabilities into the forecasting process. In this paper, we investigate this threat and present a novel attack on autoregressive diffusion models, such as those used in GenCast, capable of manipulating weather forecasts and fabricating extreme events, including hurricanes, heat waves, and intense rainfall. The attack introduces subtle perturbations into weather observations that are statistically indistinguishable from natural noise and change less than 0.1 % of the measurements—comparable to tampering with data from a single meteorological satellite. As modern forecasting integrates data from nearly one hundred satellites and many other sources operated by different countries, our findings highlight a critical security risk with the potential to cause large-scale disruptions and undermine public trust in weather forecasting.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Security and privacy** → **Software and application security**.

Keywords

Adversarial Machine Learning, Weather Forecasting, Adversarial Robustness, Security of AI

ACM Reference Format:
Erik Imgrund, Thorsten Eisenhofer, and Konrad Rieck. 2025. Adversarial Observations in Weather Forecasting. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25), October 13–17, 2025, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3719027.3765076

1 Introduction

Weather forecasting plays a central role in our daily life, ranging from choosing appropriate clothing to managing operation of renewable energy systems, agricultural planning, aviation operations, and disaster risk mitigation. In recent years, weather forecasting has seen significant advances, with AI-based approaches rapidly progressing and now beginning to surpass traditional numerical weather prediction [1, 25, 34].

Currently, the leading system in this space is *GenCast* [34], an autoregressive diffusion model developed by Google. GenCast outperforms the best traditional medium-range forecasting system, ENS [13], in both day-to-day accuracy and the prediction of extreme weather events. Due to these advances, major meteorological institutions, such as the US National Oceanic and Atmospheric Administration (NOAA) and the European Centre for Medium-Range Weather Forecasts (ECMWF), are preparing to incorporate AI-based approaches into their forecasting systems [17, 26]. With the frequency and intensity of extreme weather events increasing in recent years, this integration also represents a critical step toward more effective disaster risk mitigation on a global scale.

However, this shift also introduces a new security risk. Weather forecasting systems depend on observational data aggregated from a diverse array of organizations, each operating under different jurisdictions and guided by distinct institutional incentives [14]. Moreover, the underlying data sources are equally varied, encompassing land stations, weather balloons, aircraft, ships, and satellites [16]. This decentralized and fragmented data ecosystem creates a broad attack surface, offering adversaries multiple opportunities to tamper with observations. The potential consequences of such manipulation are severe. Reliable weather warnings, for example, are indispensable for mitigating harm by enabling timely preparation and evacuation ahead of extreme events [32].

In this paper, we explore the risk of manipulating AI-based weather forecasting systems. In particular, we introduce an attack for creating *adversarial observations*, subtle changes to measurements that mislead the predictions of a weather model. While our approach is inspired by prior techniques for generating adversarial inputs [8, 29], it addresses a key challenge specific to weather models based on autoregressive diffusion, such as GenCast. These models denoise and condition their input over multiple iterations, making standard gradient calculation technically infeasible and limiting the applicability of existing attacks. To overcome this challenge, we propose a novel approximation of the inference procedure that enables the computation of effective perturbations, capable of inducing false weather forecasts, such as fabricating non-existing extreme events or concealing real ones.

The core idea of our approach is to sample the inference process of a forecasting model at a tractable number of steps and iteratively estimate its gradient in reverse. Our approximation balances the difficulty of the attack by including both small and large noise levels which stabilizes the optimization procedure. To ensure that all changes remain inconspicuous, we apply a projection operator tailored towards weather observations, which constrains the variance of each measurement variable. As weather observations naturally exhibit variance, this projection ensures that the calculated perturbations remain indistinguishable from other sources of noise, such as measurement inaccuracies and prediction errors.

CCS '25, Taipei, Taiwan

ACM ISBN 979-8-4007-1525-9/2025/10
https://doi.org/10.1145/3719027.3765076

To analyze the efficacy of this attack, we conduct an empirical evaluation across a broad range of geographic locations and time periods, using GenCast as the target model. Specifically, we construct adversarial observations to induce extreme events at specific locations, targeting precipitation (e.g., heavy rain), wind (e.g., hurricanes), or temperature (e.g., heat waves). We observe that altering just 0.1% of the measurements is sufficient to induce false extreme events and, consequently, trigger early warning systems in practice. This fraction is smaller than that corresponding to the input from a single polar-orbiting satellite. Nearly one hundred of these satellites are currently operated by different countries with partially conflicting political interests. Furthermore, we demonstrate that an attacker can suppress actual extreme events, hindering timely preparations and potentially resulting in the loss of human lives. For example, we alter the predicted path of Hurricane Katrina (2005) to make it appear as though it would not strike New Orleans.

Our findings reveal a novel security threat that could erode trust in weather forecasting and have severe real-world consequences. As a potential defense, we investigate whether adversarial observations can be detected under theoretically ideal conditions. We find that detection success rates remain low (<3.1%), indicating that detection-based strategies are unlikely to be effective in practice. Given that certifiably robust models are not yet available for weather forecasting, we argue that large-scale deployment of AI-based weather models should be delayed unless the underlying data sources can be fully trusted.

Contributions. In summary, we make the following major contributions in this work:

- *Attack on weather forecasting.* We present the first attack targeting AI-based weather forecasting. Our attack is capable of creating adversarial observations that induce misleading predictions, such as non-existing extreme events, while remaining indistinguishable from natural noise.
- *Novel attack algorithm.* We propose a new algorithm for generating adversarial inputs for autoregressive diffusion models. The algorithm gradually approximates the inference process of weather prediction, achieving higher success rates than any existing attack.
- *Comprehensive evaluation.* We demonstrate the threat of adversarial observations by creating fake extreme events for a wide range of locations and time periods for the current best AI model GenCast. Additionally, we show that an attacker can suppress accurate extreme weather predictions.

To foster further research on the robustness of AI-based weather forecasting and to ensure the reproducibility of our experiments, we make our code and artifacts publicly available at https://github.com/mlsec-group/adversarial-observations. We also provide links to the considered weather datasets and models.

Roadmap. We provide a brief introduction to weather forecasting in Section 2 before we present our attack in Section 3. Our empirical analysis is provided in Section 4, and we investigate the detectability of the attack in Section 5. We discuss the consequences of our findings and provide recommendations in Section 6. Finally, we review related work in Section 7 and conclude in Section 8.

2 Weather Forecasting

The goal of weather forecasting is to predict future weather conditions based on past observations. In this work, we focus on global weather forecasting, which is concerned with predicting weather patterns across the entire planet. To this end, the global *weather state* of the atmosphere, $\mathbf{X} \in \mathbb{R}^{|W| \times |V|}$, is represented as a grid of nodes W distributed across the globe. Each node encodes a set of real-valued variables V corresponding to key meteorological factors, such as temperature, wind speed, and sea level pressure. By analyzing changes in the weather state over time, it becomes possible to estimate future conditions on the grid with varying degrees of confidence. Such forecasts underpin a wide range of practical applications, from predicting the output of solar and wind farms [4] to forecasting the paths of tropical cyclones [40].

Traditionally, weather forecasting has relied on *numerical weather prediction* (NWP) systems, which simulate the physical interactions between atmospheric variables to generate forecasts [11, 13]. These systems have long been the primary tool for global weather prediction. However, developing such models is highly resource-intensive and demands extensive domain expertise. Moreover, producing timely forecasts typically requires access to powerful supercomputers due to the substantial computational workload [2].

2.1 Learning-based Weather Prediction

Machine learning-based weather prediction (MLWP) has recently emerged as an alternative to traditional forecasting. Rather than simulating physical processes explicitly, these models learn from historical weather data to infer atmospheric dynamics. This allows them to capture complex relationships between variables that reflect underlying physical laws. The latest MLWP systems outperform traditional methods in both accuracy and speed, producing high-quality forecasts in under ten minutes on a single computer [1, 25, 34]. Given their effectiveness, these models are highly attractive for practical use, and different efforts are underway to integrate them into operational weather forecasting [17, 26].

GenCast[34] is currently the leading MLWP system, achieving the best performance [35] in day-to-day forecasting as well as extreme event prediction. It employs an autoregressive diffusion model to generate sequential predictions of future weather states. At its core is a denoising model d, which iteratively predicts the next state by denoising an initial estimate conditioned on the current and previous states of the global grid. This process is guided by the noise level of the initial estimate, which is gradually reduced over the denoising steps until the final prediction is obtained.

More formally, given the states $\mathbf{X}^{t-1}$ and $\mathbf{X}^{t}$, the model d generates the next predicted state $\tilde{\mathbf{X}}^{t+1} = \mathbf{Z}_n^{t+1}$ by performing n denoising steps. To this end, it begins with an initial sample $\mathbf{Z}_0^{t+1} \sim \mathcal{X}(\sigma_1)$ drawn from a noise distribution $\mathcal{X}$, parameterized by an initial noise level σ_1. Subsequently, each denoising step reduces the noise level from σ_i to σ_{i+1} according to the update rule,

$$\mathbf{Z}_{i+1}^{t+1} = d(\mathbf{X}^{t-1}, \mathbf{X}^{t}, \mathbf{Z}_i^{t+1}, \sigma_i, \sigma_{i+1}).$$

where d takes the past two states $\mathbf{X}^{t-1}$ and $\mathbf{X}^{t}$, the current estimate $\mathbf{Z}_i^{t+1}$ as well as the respective noise levels σ_i and σ_{i+1} as input. This iterative and autoregressive refinement gradually enhances prediction detail by reducing noise at each step.

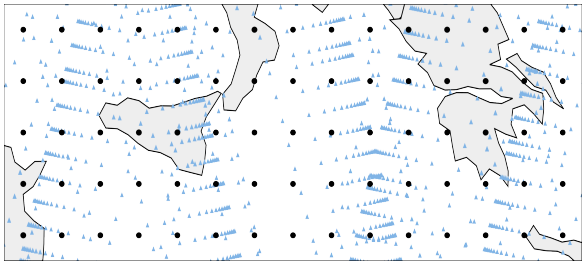


Figure 1: Location of satellite observations (blue ▲) and grid points (gray ●) for a single prediction step. The satellite paths used for reconstructing the observation locations are computed based on the orbital elements of METOP-B, METOP-C and NOAA 15 as measured by NORAD [22].

Training this model, however, poses a significant challenge: Back-propagating through all n denoising steps is computationally prohibitive. As a remedy, the model is instead trained on individual denoising steps using:

$$\tilde{\mathbf{X}}^{t+1} = d(\mathbf{X}^{t-1}, \mathbf{X}^{t}, \mathbf{Z}, \sigma, 0) \quad \mathbf{Z} \sim \mathcal{X}(\sigma), \;\; \sigma \sim \Sigma(0,1)\,,$$

where $\Sigma(a,b)$ is a probability distribution whose quantiles align with the noise levels $\sigma_1, \ldots, \sigma_n$, spanning steps $a \cdot n$ to $b \cdot n$. During training, the full noise schedule with parameters $a = 0$ and $b = 1$ is used, thereby approximating the model's behavior across all n diffusion steps.

Due to the random initialization of noise samples in each step, the prediction process is inherently stochastic. In the context of weather forecasting, this randomness is not necessarily a limitation. GenCast harnesses this stochasticity by generating multiple predictions, forming an ensemble that captures a range of plausible future scenarios. This ensemble-based approach enables uncertainty quantification and significantly enhances the so-called *forecast skill*—the ability to make accurate predictions of the global weather state [34]. Interestingly, this randomness makes constructing effective perturbations more difficult than in deterministic models.

2.2 Data Assimilation

Our discussion of weather forecasting still misses a key aspect: Real-world observations, such as temperature, pressure, and humidity, rarely align exactly with the points of a global grid. Instead, data from sources nearby the grid points must be integrated to form a consistent representation of the current weather state. This process, known as *data assimilation* in meteorology, is essential for producing accurate forecasts.

Data assimilation draws on a wide range of sources, from stationary observation points such as land stations and sea buoys to mobile platforms including balloons, aircraft, ships, and satellites [16]. While all sources are used, satellites contribute by far the largest share, providing nearly 90% of all assimilated data [15]. This dominance stems from the capability of satellites in polar orbit to scan the entire Earth's surface approximately every 12 hours [9] and geostationary satellites providing a near realtime view of a large area.

The meteorological satellites operated by several international consortia contribute to global data assimilation, such as China's CMA and NRSCC with 3 satellites, the US NOAA, NASA and US Navy with 49 satellites, and Europe's EUMETSAT and ESA with 14 satellites. As an example, Figure 1 shows measurements of three satellites within one prediction period and the respective grid points of the weather state.

Technically, data assimilation involves making an initial estimate of the current atmospheric state at a grid point, then refining it through iterative optimization [27]. This process is driven by an objective that balances two main sources of error:

- *Observation error.* This error quantifies how closely the estimated state matches actual observations. For instance, if the observed surface temperature is 20 °C but a nearby grid point predicts 0 °C, the large discrepancy results in a high observation error.
- *Background error.* This error captures the deviation between the estimated state and a short-term forecast based on previously assimilated states. The short-term forecast incorporates past observational data, thus propagating historical information into the current estimate.

The assimilated state combines current observations with short-range forecasts derived from previous states, each of which carries inherent uncertainty. To account for this uncertainty, it is explicitly modeled within the data assimilation process. This typically involves estimating the noise through the variances of observation and background errors, which are then used to regularize the assimilation procedure [16]. In contrast to the randomness in the GenCast model, this noise plays into the attacker's hand, as it allows manipulations to be concealed within the expected uncertainty of the assimilated data as we show in the following.

3 Adversarial Observations

Thus far, we have outlined how weather forecasting relies on observational data from numerous sources and is subject to inherent uncertainty in both data assimilation and inference. Building on this foundation, we now introduce our attack, which aims to manipulate forecasts by injecting adversarial observations. Before presenting the attack, we first describe the underlying threat model.

3.1 Threat Model

We characterize the threat of adversarial observations in terms of the attacker's goal, capabilities, and constraints.

Attacker's goal. We consider a scenario in which an attacker aims to manipulate forecasts generated by autoregressive diffusion models, such as those used in GenCast [34]. Potential attack goals range from causing economic harm by altering regional wind predictions, to inciting social disruption through fabricated extreme weather forecasts, and ultimately to causing physical harm by concealing impending disasters and preventing timely preparation. In this scenario, the attacker always targets a particular region instead of simply causing the general loss of predictive capabilities.

Attacker's capabilities. We assume that the adversary is capable of slightly manipulating the inputs to the forecasting model, specifically, the grid $\mathbf{X}$ assimilated from data of different meteorological sources (see Section 2). While such manipulations could, in principle, be introduced at any source, we focus on measurements from polar-orbiting satellites due to their predominance in the assimilation process—contributing over 90% [15]—and their ability to cover the entire Earth's surface within 12 hours.

Weather satellites are managed by meteorological and space agencies worldwide, including those operated by the USA, China, India, Germany, the European Union, Japan, France, and Taiwan [14]. An attacker could compromise satellite data through various means, including internal sabotage, tampering with transmissions, breaches at ground-based command centers, or by exploiting vulnerabilities within the satellite systems [33, 43]. Even more concerning, manipulations could also be deliberately introduced by an operator as part of a strategic attack against another country. Moreover, we assume that the adversary has white-box access to the forecasting model, including full knowledge of its architecture and parameters. In contrast to other domains, this assumption is plausible, as state-of-the-art learning models for weather forecasting are generally open-sourced [e.g., 1, 25, 26, 34], as no significant security concerns have been raised so far.

Attacker's constraints. We assume that any manipulation of the forecasting model's input is subject to practical constraints. For instance, control over a single satellite does not permit arbitrary modifications, as its observations are assimilated alongside data from numerous other sources. As a result, we assume that the adversary can modify only a small fraction of the values at each node in the weather state. Note that polar-orbiting satellites pass over each grid point approximately twice per day, so global perturbations are surprisingly not a limiting factor in our attack. In addition, manipulations are constrained by mechanisms designed to detect errors. Since weather forecasting is inherently imprecise, several such mechanisms are employed to reduce errors in the model's input as early as possible. Consequently, manipulations are only effective if the introduced perturbations remain within the expected variance of the input variables. In this context, adversaries can exploit the noisy nature of weather measurements but cannot introduce larger deviations without risking detection.

To model these constraints, we assume that the adversary can introduce noise with a small standard deviation, denoted by ϵ, where ϵ is smaller than the expected variance of any variable at the manipulated grid points. Furthermore, we conservatively assume that the perturbation must be unbiased, as the adversary can influence only a limited portion of the collected observational data.

3.2 Attack Methodology

Building on our threat model, we now present our attack strategy for generating adversarial observations. The core idea is to manipulate the estimated state $\tilde{\mathbf{X}}^t$ at time t so that the predicted state $\tilde{\mathbf{X}}^{t+j}$ at a later time $t+j$ aligns with a predefined target. To achieve this, the attacker adds perturbations $\boldsymbol{\delta}^t$ and $\boldsymbol{\delta}^{t-1}$ to the weather states $\mathbf{X}^t$ and $\mathbf{X}^{t-1}$, respectively, thereby influencing the prediction of $\tilde{\mathbf{X}}^{t+j}$ in the subsequent autoregressive iterations.

While $\mathbf{X}^t$ and $\mathbf{X}^{t-1}$ are unknown at attack time, the attacker can base their perturbation on the predicted states obtained using $\mathbf{X}^{t-2}$ and $\mathbf{X}^{t-3}$. More importantly, the resulting perturbations are constrained to be unbiased and limited in their variance, with standard deviations not exceeding a threshold ϵ.

Objective function. Formally, the objective can be defined through an adversarial loss function $\mathcal{A}$, which measures the distance from a selected target and is minimized by the attacker using the (approximated) inference function f of the MLWP system. In the case of GenCast, this function encapsulates the entire prediction procedure across multiple noise levels and time steps.

To model the constrained perturbations, we define a per-variable mean μ_v and standard deviation σ_v for each variable $v \in V$, which the perturbations must satisfy. These parameters allow us to constrain both the direction and the variability of the adversarial influence. Combining these elements, we arrive at the following optimization problem, where j is the lead time for the forecast and n the number of considered noise levels used within f:

$$\begin{aligned} \underset{\boldsymbol{\delta}^t, \boldsymbol{\delta}^{t-1}}{\arg\min} \quad & \mathcal{A}\left(f(\mathbf{X}^t + \boldsymbol{\delta}^t, \mathbf{X}^{t-1} + \boldsymbol{\delta}^{t-1}, j, n)\right) \\ \text{subject to} \quad & \forall v \in V : \mu_v = 0 \wedge \sigma_v \leq \epsilon. \end{aligned}$$

Decomposing the adversarial loss. The function $\mathcal{A}$ captures the complex task of manipulating forecasts in a single expression, rendering direct optimization challenging. To address this, we decompose $\mathcal{A}$ into two modular components, $\mathcal{A} = \mathcal{V} \circ \mathcal{S}$. The spatial function $\mathcal{S}$ specifies the geographic region of interest, while the variable function $\mathcal{V}$ extracts the relevant meteorological target within that region. This structured formulation provides a flexible and unified optimization framework, capable of representing a wide range of targets—from fabricating extreme winds to concealing genuine rainfall anywhere on the globe. To illustrate the utility of this decomposition, let us consider the following definition of the adversarial loss $\mathcal{A}$:

$$\begin{aligned} \mathcal{S}\colon \mathbf{X} &\mapsto \left\{\mathbf{X}_{(\text{lat},\text{lon})} \mid \text{lat} \in [51, 52], \text{lon} \in [-1, 1]\right\}, \\ \mathcal{V}\colon R &\mapsto -\min_{r \in R}\left(\sqrt{(r_{\text{u-wind}})^2 + (r_{\text{v-wind}})^2}\right). \end{aligned}$$

In this example, the spatial function $\mathcal{S}$ selects all grid points with latitudes between 51° and 52° and longitudes between −1° and 1°, corresponding to the London area. The variable function $\mathcal{V}$ then computes a scalar value from this region—specifically, negative minimum wind speed, derived from the eastward (U) and northward (V) wind components. As a result, the formulated loss function seeks to maximize the minimum predicted wind speed around London. More complex objectives can similarly be defined by customizing the spatial and variable functions.

Approximating the inference function. The diffusion model underlying f is inherently non-deterministic, as it generates forecasts by iteratively denoising samples initialized with random noise. In the case of GenCast, this process unfolds over 40 steps, making end-to-end differentiation computationally prohibitive. To mitigate this, we could adopt an approximation strategy proposed by Liang et al. [28], in which a single noise level is selected and the sample is denoised from that point onward.

However, this approximation alone does not fully resolve the challenge of determining effective perturbations for f. First, the stochastic nature of the diffusion process means that the impact of a perturbation heavily depends on the specific realization of the initial noise. Second, the influence of the denoising step varies with the selected noise level: lower noise levels result in only minor forecast changes, while higher noise levels permit more substantial alterations. As a result, the optimization process becomes highly variable, and sampling only a single the noise level ddoes not yield reliable perturbations for executing an attack.

Sampling multiple noise levels. To improve the approximation of the inference process, we introduce two key refinements, as outlined in Algorithm 1. First, rather than selecting a single noise level, we sample $n > 1$ distinct levels drawn from non-overlapping intervals across the noise distribution. Second, instead of denoising in a single step, we perform a sequence of denoising operations: the process begins with noise sampled at the first level, followed by iterative denoising through the subsequent levels, and concludes with a final denoising step from the last level to zero.

Algorithm 1: Our approximation of the autoregressive diffusion inference process.

Input: inputs $\mathbf{X}^t, \mathbf{X}^{t-1}$, lead time steps j, number of steps n
Output: approximate prediction $\tilde{\mathbf{X}}^{t+j}$

1 $\mathbf{Z}_n^t, \mathbf{Z}_n^{t-1} \leftarrow \mathbf{X}^t, \mathbf{X}^{t-1}$;
2 **for** $\tau \leftarrow t+1$ **to** $t+j$ **do**
3 Sample $\sigma_0, \ldots, \sigma_{n-1} \sim \Sigma\left(0, \frac{1}{n}\right), \ldots, \Sigma\left(\frac{n-1}{n}, 1\right)$;
4 Sample $\mathbf{Z}_0^\tau \sim \mathcal{X}(\sigma_0)$;
5 **for** $i \leftarrow 1$ **to** $n-1$ **do**
6 $\mathbf{Z}_i^\tau \leftarrow d(\mathbf{Z}_n^{\tau-2}, \mathbf{Z}_n^{\tau-1}, \mathbf{Z}_{i-1}^\tau, \sigma_{i-1}, \sigma_i)$;
7 $\mathbf{Z}_n^\tau \leftarrow d(\mathbf{Z}_n^{\tau-2}, \mathbf{Z}_n^{\tau-1}, \mathbf{Z}_{n-1}^\tau, \sigma_{n-1}, 0)$;
8 **return** $\mathbf{Z}_n^{t+j}$;

This strategy ensures that each optimization step incorporates both high and low noise levels, striking a balance between influence and difficulty. In doing so, our refined approximation more closely mimics the full inference procedure. In particular, lines 3–4 of Algorithm 1 sample from the aligned distribution Σ and the noise distribution $\mathcal{X}$ to generate an initial estimate. Subsequently, lines 5–7 iteratively refine this estimate by applying the denoising function d across a sequence of decreasing noise levels σ_i.

Projecting the perturbations. Finally, to ensure that the perturbations remain within the prescribed bounds, we introduce a projection operator Π, defined as

$$\Pi_\epsilon(\boldsymbol{\delta}) = (\boldsymbol{\delta} - \mu_v) \cdot \frac{\min(\epsilon, \sigma_v)}{\sigma_v}.$$

This operator is applied to the perturbation $\boldsymbol{\delta}$ of each variable v across all grid points. We denote this as $\Pi_\epsilon(\boldsymbol{\delta})$, indicating that the projection is performed independently for each variable. The projection ensures that the perturbations conform to the specified per-variable constraints, maintaining the prescribed zero mean and standard deviation ϵ.

Complete attack algorithm. The complete attack procedure, integrating all components and refinements, is presented in Algorithm 2. The method follows a standard gradient-based framework for generating adversarial inputs over N iterations, leveraging the approximated inference function (line 4) and applying the projection operator Π to enforce perturbation constraints (lines 6 and 9). To improve optimization efficiency, we incorporate momentum (line 6) and use a cosine annealing schedule (line 7) to dynamically adjust the step size throughout the process.

Algorithm 2: Our attack algorithm with n approximation steps of the diffusion process.

Input: inputs $\mathbf{X}^t, \mathbf{X}^{t-1}$, attack budget ϵ, steps N, lead time j
Output: adversarial perturbation $\boldsymbol{\delta}^t, \boldsymbol{\delta}^{t-1}$

1 $\mathbf{m}_0 \leftarrow \mathbf{0}$;
2 $\boldsymbol{\delta}_0 = (\boldsymbol{\delta}_0^t, \boldsymbol{\delta}_0^{t-1}) \leftarrow \mathbf{0}$;
3 **for** $i \leftarrow 1$ **to** N **do**
4 $\tilde{\mathbf{X}}^{t+j} = f(\mathbf{X}^t + \boldsymbol{\delta}_{i-1}^t, \mathbf{X}^{t-1} + \boldsymbol{\delta}_{i-1}^{t-1}, j, n)$;
5 $\mathbf{g}_i \leftarrow \nabla_{\boldsymbol{\delta}_{i-1}} \mathcal{A}(\tilde{\mathbf{X}}^{t+j})$;
6 $\mathbf{m}_i \leftarrow \beta \cdot \mathbf{m}_{i-1} + (1-\beta) \cdot \Pi_1(\mathbf{g}_i)$;
7 $\alpha_i' \leftarrow \frac{\epsilon}{N} + \frac{1}{2}\left(2\epsilon - \frac{\epsilon}{N}\right) \cdot \left(1 + \cos\left(\frac{(i-1)\cdot\pi}{N}\right)\right)$;
8 $\alpha_i \leftarrow \frac{\alpha_i'}{(1-\beta)^i}$;
9 $\boldsymbol{\delta}_i \leftarrow \Pi_\epsilon(\boldsymbol{\delta}_{i-1} - \alpha_i \mathbf{m}_i)$;
10 **return** $\boldsymbol{\delta}_N^t, \boldsymbol{\delta}_N^{t-1}$

4 Evaluation

We proceed to evaluate the effectiveness of the proposed attack in generating adversarial observations under real-world conditions. To this end, we consider two scenarios: (a) fabricating extreme events and (b) concealing extreme events. That is, we first investigate whether adversarial observations can reliably induce non-existent extreme events across various locations and points in time. Second, we examine whether the accuracy of forecasts for genuine extreme events can be compromised, for example, by moving their location or diminishing their intensity.

4.1 Experimental Setup

For all our experiments, we target GenCast [34], the currently leading MLWP system [35]. Specifically, we use the median prediction deviation from a GenCast ensemble consisting of five members and consider a one-degree grid resolution. For our attack, we generate adversarial observations two days prior to a target prediction time with $j = 4$, resulting in an attack time offset of two days. We use $N = 50$ iterative optimization steps to ensure that the resulting deviation remains robust to the stochasticity of the inference process. Additionally, we set the number of approximation steps to $n = 2$. All experiments were run on a server with four NVIDIA A40 GPUs.

Dataset. We perform all experiments using the ERA5 dataset [19], which provides hourly assimilated weather variables of the entire globe across multiple pressure levels and GenCast was trained on. For a single state $\mathbf{X}$ this amounts to approximately >5 M individual variable values across 65,160 grid points.

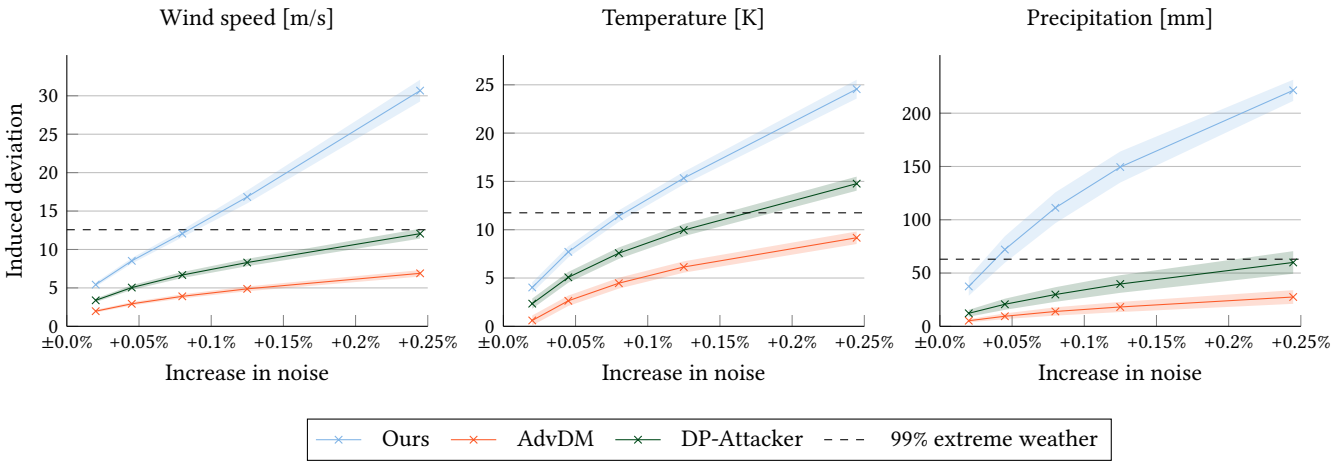


Figure 2: Resulting mean deviation induced by adversarial observations of different sizes. The average deviation of wind speed, temperature and precipitation as well as the 90% confidence interval across all target locations and times are shown. The attacker goal is to achieve the threshold for 99% extreme weather deviations with minimal noise increase.

We evaluate on data from 2022, which is the most recent full year that is publicly available as part of WeatherBench2 [35], the most common benchmark for MLWP systems.

Extreme weather. Following common practice in meteorology, we define extreme weather events based on the deviation of a target variable from its expected value. Specifically, we consider events exceeding the *99th percentile* for three variables: (a) wind speed at 10 meters above ground, (b) temperature at 2 meters above ground, and (c) precipitation accumulated over a 12-hour period. That is, we focus on wind speed, temperature, and precipitation values in the top 1 % of measurements at each location.

To determine the 99th percentile threshold for each variable, we analyze all historical weather states available in the ERA5 dataset, evaluating each target variable individually. We construct a climatological model for each variable, estimating the expected value for any given day of the year at a specific location by averaging across all available years. Using this model, we compute the maximum deviation between the expected and actual values of the variable for each year and location. We then derive the 99th percentile of these yearly maxima and average them across all grid points to obtain thresholds corresponding to the 99% extreme weather deviations.

Attacker setup. For our attack, we assume an adversary capable of manipulating data from a single polar-orbiting satellite. Under this scenario, we derive the maximum permissible standard deviation ϵ (see Section 3.1). Since the individual contribution of a single satellite cannot be precisely determined, we conservatively approximate its influence by assuming it is smaller than average: Approximately 100 meteorological satellites contribute to the ECMWF assimilation system [14], so that, on average, a single satellite accounts for roughly 1% of the total observation error. Since this error is typically larger than background error, we can set a lower bound on it using the background error [3]. Specifically, we limit the increase in noise to just 0.25 % of the standard deviation of the background error.

To map this relative constraint to absolute terms, we estimate the variance of the background error per year. As previously described, the background error is defined as the difference between the short-range forecast from the previous state and the final assimilated state. We use GenCast to perform a single-step forecast for each of our evaluation years and compute the difference to the corresponding assimilated values. Finally, we calculate the average variance across all grid points and forecasts for each variable. The resulting attack setup is conservative and clearly underestimates the potential real-world impact of compromising a single satellite.

4.2 Fabricating Extreme Events

We begin by investigating whether adversarial observations can trigger extreme weather predictions across different locations and times. To select target locations, we focus on densely populated areas. Specifically, we randomly sample 100 sites from the 1,000 most populous population centers using the Global Human Settlement Urban Centre Database R2024A [30], which provides up-to-date estimates of global population distribution based on satellite imagery analysis. The selected locations range from mid-sized cities such as Suez and Leipzig to major metropolitan areas like Los Angeles and Ho Chi Minh City. For each site, we randomly select a target time within the evaluation year 2022.

For each of these location–time pairs, we run our attack to induce extreme deviations in each of the three target weather variables at the specified location and time, manipulating the observations two days earlier ($j = 4$). To evaluate the impact of perturbation strength, we conduct the attack using logarithmically spaced noise budgets, starting from 0.02 % and increasing up to the derived maximum of 0.25 %, as discussed in Section 4.1.

Attack performance. Figure 2 displays the resulting deviations across all target variables. We observe that adversarial observations consistently induce substantial changes in weather predictions. For each of the three target variables, the noise required to fabricate extreme weather remains below the maximum allowed perturbation.

On average, triggering extreme weather conditions for temperature and wind speed requires a noise level of approximately 0.08 %, while precipitation proves even more sensitive, with the threshold for extreme weather surpassed at noise levels below 0.05 %. To put these numbers into perspective, at the maximum permitted noise level of 0.25 %, the attack can increase wind speeds by 30.7 m/s—equivalent to 111 km/h—*averaged* over a 12-hour period. This average is on par with peak wind speeds typically observed during a Category 1 hurricane. Similarly, temperatures can be increased by 24.6 C, while precipitation can be increased by 221 mm over a 12-hour period—equivalent to 221 l/m^2. This level of rainfall is comparable to that seen during extreme storm events. These results demonstrate that even minimal perturbations to observations can lead to substantial shifts in forecast outputs, highlighting the vulnerability of state-of-the-art weather prediction systems to adversarial manipulation.

Baselines. Next, we consider the performance of our attack against two recently proposed methods targeting diffusion models. The first, AdvDM [28], introduces perturbations directly into the noise used by image diffusion models. The second, DP-Attacker [8], is a more recent method designed to target policy diffusion models. These models generate multi-step policies autoregressively from an initial vision input, which is more similar to weather prediction and makes this attack naturally suited to our context. Both baseline attacks operate using a single sampled noise level for prediction, consistent with the noise sampling employed during training.

The results are included in Figure 2. Our attack consistently outperforms the baseline methods across all target variables and attack budgets. Notably, both baselines fail to reach the extreme weather thresholds for wind speed and precipitation. Only DP-Attacker achieves the temperature threshold with an attack budget below the maximum noise increase of 0.25 %. When comparing the baselines with our method, we observe that the performance gap widens as the attack budget increases. This suggests that our approach scales more efficiently as the budget grows. This advantage is particularly evident in the case of precipitation, where our method surpasses the baselines by a substantially larger margin.

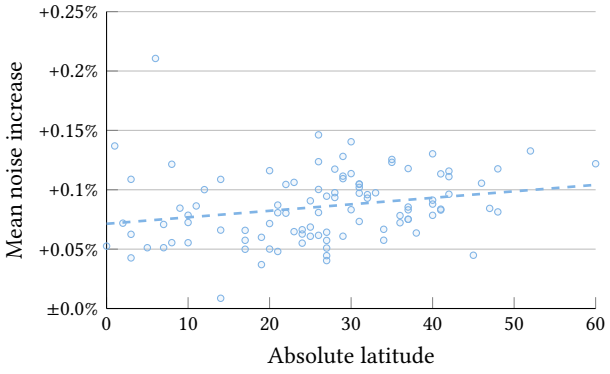


Figure 3: Mean required noise increase at different locations. The dashed line shows a linear regression of the required noise. The mean increase in noise required to fabricate an extreme weather prediction grows with increasing distance from the equator.

Table 1: Mean relative deviation achieved by different ablations. The deviation is relative to the original attack and averaged across 200 different target combinations.

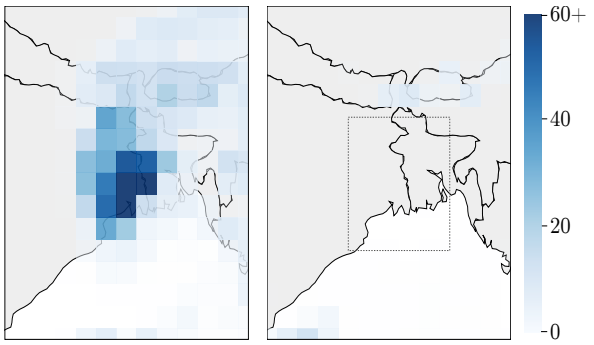
Method	Wind Speed	Temperature	Precipitation
Ours	100.0 %	100.0 %	100.0 %
w/o steps	89.3 % (-10.7)	93.1 % (- 6.9)	54.4 % (−45.6)
w/o approx	59.3 % (-40.7)	71.6 % (-28.4)	33.9 % (−66.1)
w/o both	56.0 % (-44.0)	62.9 % (-37.1)	18.4 % (−81.6)

Susceptibility of different locations. To explore how the choice of target location influences the attack, we investigate whether predictions at certain locations are more susceptible to adversarial observations than others. This is evaluated by calculating the average increase in noise required at each target location to achieve extreme weather, averaged over all target variables.

Our findings, illustrated in Figure 3, indicate a relationship between the required noise and the angular distance from the equator. Specifically, locations farther from the equator tend to require more noise to achieve the same level of deviation ($p < 0.05$). Still, even the most impacted areas require less than the maximum noise increase to trigger extreme weather predictions—indicating that, although the effect is statistically significant, its practical impact is modest. We hypothesize that this trend is linked to the uneven distribution of grid points near the equator. Because the grid is constructed with uniform spacing in both latitude and longitude, grid points become increasingly dense toward the poles and sparser near the equator. One potential solution to address this imbalance is to use a mesh derived from an icosahedron, which ensures uniform spacing between grid points regardless of geographic location. This approach aligns well with existing infrastructure, as GenCast already employs a six-times refined icosahedral grid internally. However, this adjustment alone does not resolve the underlying vulnerability.

Ablation study. To better understand the contribution of individual components within our attack methodology, we perform an ablation study. Specifically, we evaluate three simplified variants: (1) replacing our improved approximation of the inference process with the naïve approach used during training, (2) removing optimization enhancements such as cosine annealing and momentum, and (3) removing both simultaneously. Due to computational constraints, we restrict our evaluation to a subset of 200 out of the original 1,500 target combinations. For each variant, we compute the prediction deviation relative to the full attack, quantifying the extent to which each component contributes to overall effectiveness.

The results in Table 1 show removing any single component leads to a noticeable drop in performance. For temperature and wind speed, most of the performance is retained without the optimization improvements. This suggests that the inference approximation plays a more critical role for these variables. Moreover, removing the inference approximation alone has a larger impact than removing the improved steps. As expected, disabling both components results in the largest performance reduction. These findings suggest that the interplay between the inference approximation and the optimization enhancements is essential to achieving the strong attack effectiveness observed in our earlier experiments.



(a) Original prediction **(b)** Perturbed prediction

Figure 4: Predicted precipitation at the peak of Cyclone Amphan. The forecast is shown (a) without attack and (b) after including adversarial observations. The dashed rectangle depicts the target region of the attack. The precipitation is expressed as mm over a 12-hour period.

4.3 Concealing Extreme Predictions

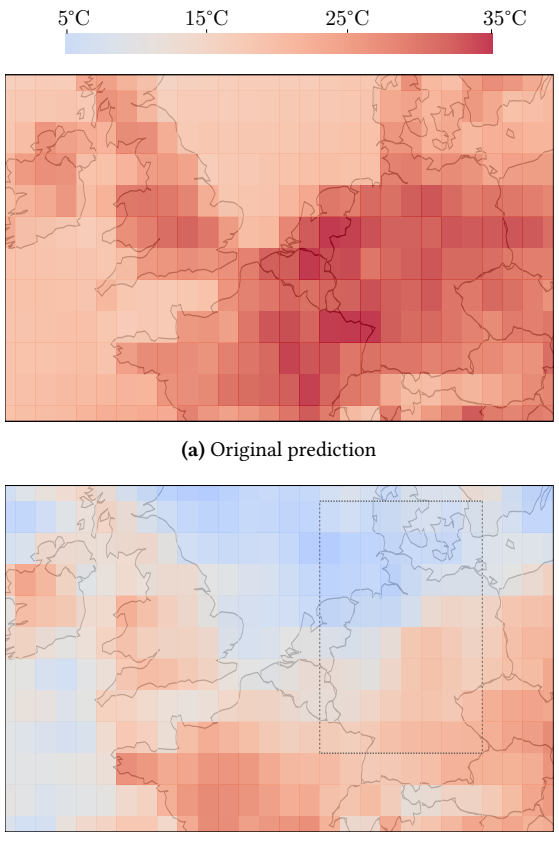
Thus far, our analysis has focused on scenarios in which an adversary seeks to fabricate predictions of extreme weather at specific times and locations. We now turn to a different question: can the attack also undermine genuine forecasts of extreme weather events? To explore this, we apply our method to three major historical events—Cyclone Amphan (2020), the 2006 European heat wave, and Hurricane Katrina (2005). For each event, we simulate an attack by introducing a maximum noise perturbation of $\epsilon \leq 0.25\,\%$ into the weather state two and a half days before the event reaches peak intensity. This time frame ensures that the extent and location of the extreme event are correctly predicted without the attack but also allows for realistic manipulation. Differing from the previous section, the attacker's goal is not to increase the intensity at a single grid point but instead to reduce it in a region.

Cyclone Amphan. Our first case study focuses on tropical Cyclone Amphan, which struck Bangladesh, India, and Sri Lanka in May 2020, bringing strong winds and heavy rainfall that caused widespread flooding [23]. Several days prior to landfall, the storm significantly intensified—which was correctly predicted by GenCast—leading up to intense precipitation across the region as shown in Figure 4a as the blue shaded area.

We adversarially perturb the observations before this intensification, targeting a prediction outcome with minimal precipitation across the expected storm region. As shown in Figure 4b, the resulting forecast entirely suppresses precipitation in the target region. Notably, when examining the sequence of predicted states between the perturbed inputs and the forecast, we observe a plausible dissipation of the storm. In this manipulated scenario, the storm releases rainfall over the ocean and weakens before reaching land. This illustrates how an attacker could convincingly mask an otherwise accurate forecast of a severe weather event. Crucially, the perturbations as well as the intermediate weather development appear plausible despite the underlying manipulation.

European Heat Wave. To assess the ability of our attack to conceal extreme temperatures, we apply it to the European Heat Wave of 2006, which set temperature records across a multitude of Western European countries [36]. Figure 5 presents the temperature forecasts before and after the introduction of adversarial observations. As in the previous case study, the extreme weather signal is effectively suppressed in the targeted region following the attack.

For this specific attack, we include only the eastern portion of the heat wave as the target region (indicated by the rectangle in Figure 5). Despite this narrow focus, extreme temperatures are also eliminated from adjacent areas. This highlights another key insight: the impact of adversarial observations extends beyond the targeted geographic region, plausibly removing the entire extreme weather event rather than confining the effect locally. Furthermore, we observe that the altered forecast significantly overshoots the intended objective of merely hiding the heat wave. Instead, unnaturally mild temperatures ranging from 5°C to 10°C in the regions surrounding the North Sea are predicted. This effect could be mitigated by specifying a desired target temperature, instead of naïvely minimizing the predicted temperature.



(a) Original prediction

(b) Perturbed prediction

Figure 5: Predicted temperature at the peak of the European Heat Wave 2006. The forecast is shown (a) without attack and (b) after including adversarial observations. The dashed rectangle depicts the target region of the attack.

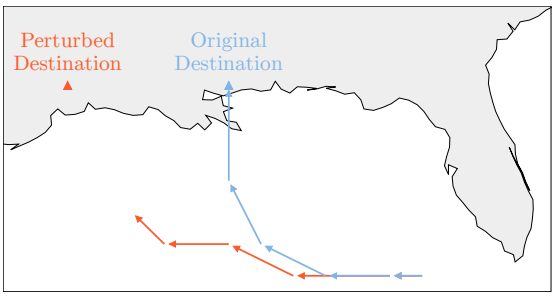


Figure 6: Predicted storm path of Hurricane Katrina. The forecast is shown (a) without attack and (b) after including adversarial observations. The triangles show the target location at which the wind speed is minimized (▲) and maximized (▲).

Hurricane Katrina. Our final case study evaluates the precision with which an adversary manipulate the course of a storm, using Hurricane Katrina as an example. After initially passing Florida, the storm made its primary landfall near New Orleans [42]. Rather than suppressing the storm entirely, an adversary may aim to shift the predicted landfall site to disrupt relevant preparations. To simulate this, we compute adversarial observations that reduce the predicted wind speed at the original landfall site while simultaneously increasing it at a new, perturbed location.

The original and perturbed storm tracks are shown in Figure 6. After introducing the perturbation, the forecast storm path clearly deviates from the original, no longer indicating landfall near New Orleans but instead pointing to the manipulated location. The storms trajectory is determined using the location of lowest sea level pressure, which serves as a proxy for the storms eye. Notably, although the optimization process targets wind speed predictions, it also affects atmospheric pressure, again suggesting broader implications of adversarial interference.

5 Statistical Detection

Our findings demonstrate that AI-based weather forecasting systems are vulnerable to adversarial observations, underscoring the need for effective defense mechanisms. In the following, we thus explore statistical detection as a potential countermeasure to mitigate this vulnerability, while broader organizational responses are discussed in Section 6.

Noisy data is not unique to the adversarial context and in fact a common and practical challenge for real-world forecasting systems. To manage this, quality control procedures are implemented that evaluate the reliability and plausibility of incoming data. These procedures typically consist of hand-crafted rules involving two main categories: whether the observations are temporally and spatially consistent, and whether they fall within a reasonable range of the best estimate of the value [12, 41]. Because these checks are designed to handle naturally occurring noise and errors, they are insufficient for detecting the subtle, worst-case perturbations introduced by adversarial observations. We therefore explore whether more sophisticated statistical tests could identify manipulations and serve as a defense against this threat.

We evaluate detecting adversarial observations in the context of a statistical difference to real data. The assimilated state $\bar{\mathbf{X}}$ is commonly assumed to consist of an unknown underlying ground-truth value $\mathbf{X}$, to which unbiased Gaussian noise is added by the background and observation error [3]. We assume a best-case scenario for the defender in which all natural noise can be described by the background error alone. In this setting, the attacker adds noise through the adversarial observations and we arrive at

$$\bar{\mathbf{X}} = \mathbf{X} + \mathcal{N}(0, \sigma_b^2) + \mathcal{N}(0, \epsilon^2) = \mathbf{X} + \mathcal{N}\left(0, \sigma_b^2 + \epsilon^2\right),$$

where σ_b^2 denotes the variance of the background error.

Under this formulation, any adversarial perturbation increases the total noise in the assimilated state. Thus, if the background error variance is both constant and known exactly, the presence of an attack can, in principle, always be detected—provided the sample size is sufficiently large—since the resulting variance will exhibit a measurable increase. In practice, however, the sample size is constrained by the number of grid points and the number of variables per grid point, making detection inherently probabilistic. Moreover, the variances of background and observation errors are neither constant nor known with high precision, which makes detecting small increases in noise particularly challenging.

Despite these limitations, we take a conservative approach to evaluate the overall detectability of the attack, assuming a best-case scenario for the defender in which the total error variance in the assimilated state is constant and known. Under this assumption, we can determine whether a given sample shows a significantly higher variance by applying a simple chi-square test for the variance [31].

Chi-square test setup. We consider the targets described in Section 4 and compute the minimum increase in noise required to trigger an extreme weather deviation. This is estimated by linearly interpolating the induced deviations across the evaluated noise levels. To ensure that each attack can reach the extreme weather threshold, we do not impose a limit on the maximum noise level. In such cases, we extrapolate beyond the defined maximum attack budget. For all attacks, we then estimate the detection probability using a chi-square test for variance, assuming perfect knowledge of the expected amount of noise.

Detection results. The detection probabilities are presented in Table 2. Adversarial observations from both baselines are consistently detected using the chi-square test, with rates exceeding 95 % in all cases—except when temperature is manipulated by the DP-Attacker. In contrast, our attack results in significantly lower detection probabilities: approximately $\approx 3\%$ for wind speed and temperature, and just 0.2 % for precipitation.

These results demonstrate that, even under ideal conditions, the attack would likely evade detection. This conclusion is further reinforced by the fact that the assumed detection method is not practically feasible and would likely result in false positives. Consequently, even if such a method were implementable, successful detection would remain unlikely and establishing definitive proof of an attack even more so. We therefore conclude that statistical detection is, unfortunately, not a viable approach for defending against adversarial perturbations in weather forecasting.

Table 2: Detectability of different approaches used to fabricate extreme weather deviations. The detectability is measured using a chi square test for the variance with best-case assumptions of constant and perfectly known variance of the assimilation error.

Method	Wind Speed	Temperature	Precipitation
AdvDM	> 99.99 %	99.92 %	> 99.99 %
DP-Attacker	95.04 %	45.85%	95.33 %
Ours	3.07 %	2.96 %	0.20 %

6 Discussion

Our findings highlight a critical vulnerability in modern weather forecasting: the integration of machine learning into the prediction pipeline introduces a new attack surface for manipulation. These concerns align with prior research that has revealed fundamental limitations in the robustness of machine learning systems [24, 29]. Even more concerning, as demonstrated in Section 5, such manipulations are likely to remain undetected. While crafting adversarial observations may exceed the capabilities of typical cybercriminals, they represent a promising tool for more sophisticated and well-resourced actors, including nation-state adversaries. In the following, we thus take a broader perspective on the impact of our work, beginning with a discussion of its limitations and assumptions, followed by recommendations for mitigating this threat.

6.1 Limitations

We begin by outlining the key assumptions underlying our attack and how they may limit its practical impact.

Access to prediction model. Our attack relies on computing gradients of the model's outputs, which requires access to the model weights. Currently, this is a reasonable assumption, as many leading forecasting models are publicly available [e.g., 1, 25, 26, 34]. However, it is possible that future models will not be publicly released, which would significantly hinder an adversary's ability to carry out the attack. Black-box attacks on machine learning models typically require vastly more queries to the target model [5], rendering such approaches impractical for weather models. This difficulty is further exacerbated by the operational nature of weather forecasting systems, which generally produce predictions only once per time step. For example, conducting 1,000 queries—on the very low end of what is typical for black-box attacks—against a model with a 12-hour time step would require 500 days to complete.

To overcome this constraint, black-box methods would likely need to identify a universal adversarial perturbation that remains effective across multiple time steps. A more feasible alternative arises if the attacker has regular access to the output forecasts of the target system. In this case, a model extraction (or model stealing) attack could be performed, allowing the adversary to reconstruct an approximate surrogate of the target model over time. Adversarial observations could then be crafted using this surrogate in a white-box setting and transferred to the original system. However, model extraction would be slow in this case, as the attacker cannot control the inputs, and thus the process would again require a significant amount of time.

Continuous attack. In this work, we focus to introduces perturbations at a time step t to manipulate the prediction at a future time step $t + j$. In practice, however, weather forecasts are updated continuously and new predictions are typically made at each time step. This would require the attack to sustain the manipulations until reaching the forecast at $t + j$. This adds a layer of complexity to the attack, as the adversary must persist with the attack long enough for decisions to be influenced by the forecast. This persistence does not have to be negative and could also work in the adversary's favor. Since data assimilation implicitly incorporates the entire history of observations, it may be possible to exploit this process, potentially making the attack more effective. Furthermore, if we extend our view to earlier time steps, smaller adversarial perturbations could be distributed over a longer period, potentially making the attack more subtle and harder to detect.

Weather forecasting pipeline. We consider an attack on weather prediction, which is a single step of the weather forecasting pipeline. In particular, on the assimilated state on the grid, while an attacker can only control the observations before data assimilation and with the output of the weather prediction being the target without any further statistical postprocessing. Although this might seem to constrain our attack to the realm of theoretical feature-space attacks, we have ensured a practical scenario by considering the problem space across all points. The influence of the attacker is realistic in adding only noise and the derived constraint is both conservative and faithful to real-world constraints. Furthermore, our statistical detection is not only inspired by real-world quality control procedures, but assumes a far stronger defender that still cannot reliably detect our attack. Additionally, current developments indicate that data assimilation will also be integrated to achieve end-to-end AI-based weather forecasting in the near future [1, 21]. This would enable directly computing gradients to the individual observations, allowing attackers to perform the same attack directly on the problem space. Moreover, further statistical postprocessing is unlikely to nullify the effects of our attack, because it would be based on the compromised outputs of the weather prediction.

6.2 Countermeasures

Given the limitations of detecting adversarial observations, we consider alternative defense strategies that extend beyond detection.

Hybrid methods. A straightforward approach to enhancing forecasting robustness is to cross-verify predictions using traditional numerical weather prediction (NWP) systems whenever extreme weather events are forecast. This strategy preserves the benefits of shorter runtimes and improved accuracy offered by the MLWP system, while potentially mitigating exposure to adversarial threats.

However, this approach is not sufficient. First, such selective verification would fail to detect the second attack scenario, where an adversary suppresses an impending extreme weather event from the forecast since no secondary check would be triggered. Moreover, even for the first attack scenario, a conflicting forecast from NWP would not prove an attack or an error of MLWP, given that MLWP has demonstrated the ability to predict extreme events earlier with greater accuracy [34]. Consequently, operating hybrid MLWP and NWP systems does not constitute an adequate countermeasure.

Adversarial robustness. In other domains, adversarial training has proven effective in improving the robustness of machine learning models [20, 29]. However, given the complexity and immense computational resources required to train state-of-the-art forecasting models, adversarial training is likely prohibitively expensive or negatively impacts performance relative to traditional prediction systems. We leave a deeper exploration of this approach to future work within the meteorological community. As a more practical remedy, we recommend that future MLWP development prioritize not only forecast accuracy but also robustness, by systematically evaluating models against attacks, such as ours. While this strategy may not entirely eliminate the risk of adversarial observations, it could raise the noise threshold required for a successful attack, reducing its impact or making it more likely to be detected through statistical testing.

Better detection. Even though we demonstrate that statistical detection is unlikely even in the best case when modeling the background and observation errors as Gaussian noise, as is common in meteorology, other models might allow better detection. For example, detection of physical inconsistencies could indicate a perturbation, as the attack does not currently incorporate strict physical dependencies into the constraints. Of course, those constraints also are not fully held for naturally occurring observations because of measurement errors. Anomaly detection of more powerful statistical tests might provide further avenues for improved detection but could also fall into the trap that extreme weather conditions are anomalous by definition. Thus, detection methods always introduce a trade off between better robustness against attacks and the chance of false positives.

Trusted data sources. The existence of adversarial observations highlights a fundamental dependency on the integrity of data sources used in weather forecasting. In safety-critical contexts—such as military or space operations—this dependency necessitates the exclusive use of trusted and rigorously validated observational inputs. Although this constraint may reduce forecast accuracy, it significantly lowers the risk posed by adversarial data. However, such measures cannot entirely eliminate the threat, as a determined adversary may still succeed in compromising individual sources without detection by any trusted entity.

7 Related Work

A substantial body of research has focused on generating adversarial examples for machine learning classifiers [6, 10, 24, 29]. In contrast, comparatively little attention has been devoted to attacks targeting diffusion models or weather forecasting systems.

7.1 Attacks on Diffusion Models

Diffusion models were initially developed and explored in the image domain, where they also faced the first wave of attacks. Initial efforts focused on identifying perturbations that make images unlearnable, aiming to safeguard intellectual property [38]. Subsequent work explored how adversarial examples could be used to prevent imitation or replication of specific artistic styles in generated images. An example is AdvDM by Liang et al. [28] that we consider in our evaluation.

The diffusion models targeted by these approaches, however, differ significantly from those used in weather forecasting. In image generation, the models typically denoise a target sample directly, whereas in weather forecasting, they generate sequences of samples autoregressively.

More recently, diffusion models have been extended to domains such as robotic control, which more closely parallels weather forecasting due to its reliance on autoregressive sampling. Attacks in this domain have emerged as well, crafting inputs that disrupt a robot's ability to complete its tasks. Despite domain-specific variations, the core attack strategies are largely consistent, generally relying on single-step denoising to approximate the inference process. We consider the approach DP-Attacker by Chen et al. [8] in our evaluation. However, our findings reveal that such attacks fail to produce adversarial observations with perturbations small enough to be considered imperceptible or practically effective.

7.2 Attacks on Weather Forecasting

Attacks have also been explored in the context of weather forecasting, specially for renewable energy planning [7, 18, 37, 39]. These studies differ significantly from ours in terms of their threat model. Specifically, they assume that the adversary has direct access to manipulate either the outputs of the forecasting system or the historical data of renewable energy generation. In contrast, we adopt a more realistic and plausible threat model, where adversarial perturbations are introduced through corrupted observations by a malicious actor. Moreover, prior works are limited in both scope and objective, each focusing on a single forecasting goal within a localized region. Our approach, by comparison, evaluates a broader set of attacker goals spanning global locations.

8 Conclusion

AI-based weather forecasting has attracted increasing attention, with leading meteorological institutions actively exploring the integration of such models into operational forecasting systems. Yet, despite notable advances in model architecture and performance, existing systems lack safeguards against adversarial manipulation of input data. In this paper, we demonstrate that diffusion models—such as those used in GenCast—are susceptible to precisely crafted adversarial observations that can alter extreme weather forecasts without significantly affecting the statistical properties of the input. More broadly, we introduce a novel attack framework for generating adversarial examples targeting autoregressive diffusion models, designed to operate under realistic constraints.

Responsible Disclosure

We have initiated a responsible disclosure process with the GenCast development team. We hope to explore new countermeasures in cooperation with the developers.

Acknowledgments

This work was supported by the German Federal Ministry of Research, Technology and Space under the grant AIgenCY (16KIS2012) and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project ALISON (492020528).

References

[1] Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P. Bruinsma, Tom R. Andersson, Michael Herzog, Nicholas D. Lane, Matthew Chantry, J. Scott Hosking, and Richard E. Turner. 2025. End-to-End Data-Driven Weather Prediction. *Nature* (2025).

[2] Peter Bauer, Alan Thorpe, and Gilbert Brunet. 2015. The Quiet Revolution of Numerical Weather Prediction. *Nature* (2015).

[3] Niels Bormann. 2015. Observation Errors. Lecture Slides.

[4] Sebastian B. M. Bosma and Negar Nazari. 2022. Estimating Solar and Wind Power Production Using Computer Vision Deep Learning Techniques on Weather Maps. *Energy Technology* (2022).

[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2018. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *Proc. of International Conference on Learning Representations (ICLR).*

[6] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *Proc. of the IEEE Symposium on Security and Privacy (S&P).*

[7] Yize Chen, Yushi Tan, and Baosen Zhang. 2019. Exploiting Vulnerabilities of Load Forecasting Through Adversarial Attacks. In *Proc. of the ACM International Conference on Future Energy Systems.*

[8] Yipu Chen, Haotian Xue, and Yongxin Chen. 2024. Diffusion Policy Attacker: Crafting Adversarial Attacks for Diffusion-based Policies. In *Proc. of the Conference on Neural Information Processing Systems (NeurIPS).*

[9] Cathy Clerbaux, Anne Boynard, Lieven Clarisse, Maya George, Juliette Hadji-Lazaro H. Herbin, Daniel Hurtmans, Matthieu Pommier, Ariane Razavi, Solène Turquety, Catherine Wespes, and Pierre-François Coheur. 2009. Monitoring of Atmospheric Composition Using the Thermal Infrared IASI/MetOp Sounder. *Atmospheric Chemistry and Physics* (2009).

[10] Francesco Croce and Matthias Hein. 2020. Reliable Evaluation of Adversarial Robustness With an Ensemble of Diverse Parameter-Free Attacks. In *Proc. of the International Conference on Machine Learning (ICML).*

[11] European Centre for Medium-Range Weather Forecasts (ECMWF). 2024. *IFS Documentation CY49R1.* European Centre for Medium-Range Weather Forecasts.

[12] European Centre for Medium-Range Weather Forecasts (ECMWF). 2024. *IFS Documentation CY49R1 - Part I: Observations.* European Centre for Medium-Range Weather Forecasts.

[13] European Centre for Medium-Range Weather Forecasts (ECMWF). 2024. *IFS Documentation CY49R1 - Part V: Ensemble Prediction System.* European Centre for Medium-Range Weather Forecasts.

[14] European Centre for Medium-Range Weather Forecasts (ECMWF). 2025. *ERA5: Data Documentation - Observations.*

[15] European Centre for Medium-Range Weather Forecasts (ECMWF). 2025. *Section 2.4 Atmospheric Model Data Sources.*

[16] John Eyre, William Bell, James Cotton, Stephen English, Mary Forsythe, Sean Healy, and Edward Pavelin. 2022. Assimilation of Satellite Data in Numerical Weather Prediction. Part II: Recent Years. *Quarterly Journal of the Royal Meteorological Society* (2022).

[17] Sergey Frolov, Kevin Garrett, Isidora Jankov, Daryl Kleist, Jebb Q. Stewart, and John Ten Hoeve. 2024. Integration of Emerging Data-Driven Models into the NOAA Research-to-Operations Pipeline for Numerical Weather Prediction. *Bulletin of the American Meteorological Society* (2024).

[18] René Heinrich, Christoph Scholz, Stephan Vogt, and Malte Lehna. 2024. Targeted Adversarial Attacks on Wind Power Forecasts. *Machine Learning* (2024).

[19] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. 2020. The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society* (2020).

[20] Ashish Hooda, Neal Mangaokar, Ryan Feng, Kassem Fawaz, Somesh Jha, and Atul Prakash. 2023. Theoretically Principled Trade-off for Stateful Defenses against Query-Based Black-Box Attacks. *Computing Research Repository* (2023).

[21] Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter D. Düben, and Torsten Hoefler. 2024. DiffDA: A Diffusion Model for Weather-Scale Data Assimilation. In *Proc. of the International Conference on Machine Learning (ICML).*

[22] Thomas Sean Kelso. 2025. CelesTrak NORAD General Perturbations Element Sets. Online Data Collection.

[23] Shubham Kumar, Preet Lal, and Amit Kumar. 2021. Influence of Super Cyclone "Amphan" in the Indian Subcontinent amid COVID-19 Pandemic. *Remote Sensing in Earth Systems Sciences* (2021).

[24] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial Examples in the Physical World. In *Proc. of the International Conference on Learning Representations (ICLR).*

[25] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. 2023. Learning Skillful Medium-Range Global Weather Forecasting. *Science* (2023).

[26] Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, et al. 2024. AIFS–ECMWF's Data-Driven Forecasting System. *Computing Research Repository* (2024).

[27] François-Xavier Le Dimet and Olivier Talagrand. 1986. Variational Algorithms for Analysis and Assimilation of Meteorological Observations: Theoretical Aspects. *Tellus A: Dynamic Meteorology and Oceanography* (1986).

[28] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. 2023. Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples. In *Proc. of the International Conference on Machine Learning (ICML).*

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proc. of the International Conference on Learning Representations (ICLR).*

[30] Ines Marí Rivero, Michele Melchiorri, Pietro Florio, Marcello Schiavina, Katarzyna Krasnodębska, Panagiotis Politis, Johannes Uhl, Martino Pesaresi, Luca Maffenini, Patrizia Sulis, Monica Crippa, Diego Guizzardi, Enrico Pisoni, Claudio Belis, Jacome Felix Oom Duarte, Alfredo Branco, E. an Njagi Moses Mwaniki Kochulem, Daniel Githira, Pierpaolo Tommasi, Allesandra Carioli, Daniele Ehrlich, Thomas. Kemper, and Lewis Dijkstra. 2024. GHS Urban Centre Database, Multitemporal and Multidimensional Attributes, R2024A. European Commission, Joint Research Centre (JRC).

[31] NIST. 2012. *Engineering Statistics Handbook.*

[32] Florian Pappenberger, Hannah L Cloke, Dennis J Parker, Fredrik Wetterhall, David S Richardson, and Jutta Thielen. 2015. The Monetary Benefit of Early Flood Warnings in Europe. *Environmental Science & Policy* (2015).

[33] PistonMiner. 2024. Hacking Yourself a Satellite - Recovering BEESAT-1. Hacking Conference Talk.

[34] Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter W. Battaglia, Rémi R. Lam, and Matthew Willson. 2025. Probabilistic Weather Forecasting with Machine Learning. *Nature* (2025).

[35] Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russel, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. 2023. WeatherBench 2: A Benchmark for the Next Generation of Data-Driven Global Weather Models. (2023).

[36] M. Rebetez, O. Dupont, and M. Giroud. 2009. An Analysis of the July 2006 Heatwave Extent in Europe Compared to the Record Year of 2003. *Theoretical and Applied Climatology* (2009).

[37] Everton Jose Santana, Ricardo Petri Silva, Bruno Bogaz Zarpelão, and Sylvio Barbon Junior. 2021. Detecting and Mitigating Adversarial Examples in Regression Tasks: A Photovoltaic Power Generation Forecasting Case Study. *Information* (2021).

[38] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. In *Proc. of the USENIX Security Symposium.*

[39] Ningkai Tang, Shiwen Mao, and R. Mark Nelms. 2021. Adversarial Attacks to Solar Power Forecast. In *Proc. of IEEE Global Communications Conference (GLOBECOM).*

[40] Paul A Ullrich, Colin M Zarzycki, Elizabeth E McClenny, Marielle C Pinheiro, Alyssa M Stansfield, and Kevin A Reed. 2021. TempestExtremes v2.1: A Community Framework for Feature Detection, Tracking, and Analysis in Large Datasets. *Geoscientific Model Development* (2021).

[41] United States Department of Commerce National Oceanic and Atmospheric Administration National Weather Service Office of Systems Development. 1994. *Technique Specification Package 88-21-R2.* Technical Report.

[42] Jacob Vigdor. 2008. The Economic Aftermath of Hurricane Katrina. *Journal of Economic Perspectives* (2008).

[43] Johannes Willbold, Moritz Schloegel, Manuel Vögele, Maximilian Gerhardt, Thorsten Holz, and Ali Abbasi. 2023. Space Odyssey: An Experimental Software Security Analysis of Satellites. In *Proc. of the IEEE Symposium on Security and Privacy (S&P).*