# Assessing the Potential of Generative Models for Solar Flare Forecasting

Zijie Tian[1], **Kevin Jin**[2], Yang Chen[2], Thomas Lee[1]

[1]University of California, Davis, [2]University of Michigan

August 3, 2026

Section 1

# Background

# Solar Flare Intensity is measured by Soft X-ray Flux Values

The GOES X-ray Sensor measures the solar irradiance at Earth in the $0.1$–$0.8\,\mathrm{nm}$ ($1$–$8\,\text{Å}$) band and are used to track solar activity and solar flares, as large solar flares can change the Earths ionosphere, causing many negative downstream effects like loss of radio transmissions and dangerous CME events.

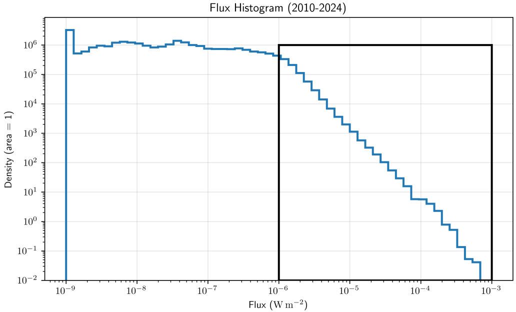



Figure: Soft X-ray Flux Histogram from 2010-2024. Box highlights the threshold of M-class and stronger flares: $10^{-6}\mathrm{W}\cdot\mathrm{m}^{-2}$

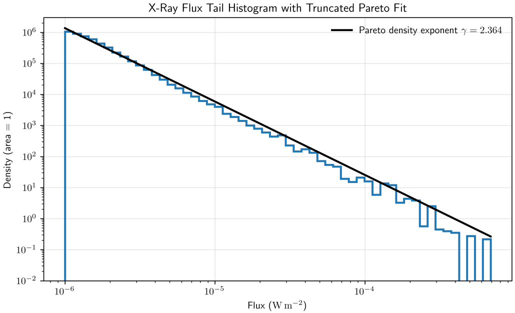



Figure: Truncated Soft X-ray Flux Histogram with Power Law Fit: $\gamma = 2.364$

# Generative Models: Diffusion Models and Flow-Matching Generators

Modern continuous-time generative models transform samples from a simple reference distribution into samples from a potentially complex conditional data distribution:

$$\boldsymbol{X}_0 \sim p_0 \quad \longrightarrow \quad \boldsymbol{X}_t \sim p_t(\,\cdot \mid y) \quad \longrightarrow \quad \boldsymbol{X}_1 \sim p_{\mathrm{data}}(\,\cdot \mid y).$$

### Score-based diffusion models (Song et al., 2020)

A forward stochastic process gradually adds noise:

$$\mathrm{d}\boldsymbol{X}_t = f_t(\boldsymbol{X}_t)\,\mathrm{d}t + g_t\,\mathrm{d}\boldsymbol{W}_t.$$

A neural network estimates the score of each noisy marginal via Tweedie's:

$$\boldsymbol{s}_\theta(\boldsymbol{x}, t, y) \approx \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x} \mid y).$$

Samples are generated by solving the reverse-time SDE or probability-flow ODE from noise to data.

### Flow-matching models (Tong et al., 2024)

FM learns a time-dependent velocity field:

$$\boldsymbol{v}_\theta(\boldsymbol{x}, t, y) \approx \boldsymbol{u}_t(\boldsymbol{x} \mid y),$$

where $\boldsymbol{u}_t$ transports the reference distribution along a chosen probability path.
Generation uses the deterministic ODE

$$\frac{\mathrm{d}\boldsymbol{X}_t}{\mathrm{d}t} = \boldsymbol{v}_\theta(\boldsymbol{X}_t, t, y), \qquad \boldsymbol{X}_0 \sim p_0.$$

Optimal-transport couplings can define relatively direct paths between noise and observed data.

# Synthetic Data are not Automatically Useful

- Real data comes from some target law $p(x, y)$, but generated synthetic data can come from a different proposal law $q_{\mathrm{prop}}(x, y)$.
- The proposal may be intentionally different from the target: rare-event enrichment, coverage of underrepresented regions, simulator bias, or generator-induced mismatch.
- Importance weighting corrects expectations, but can increase variance when density ratios or weighted scores are unstable.

### Acceptance-design Question

Given a large candidate synthetic pool, which samples should enter the final weighted estimator?

Section 2

# `REAL-Syn` Methodology

# REAL-Syn

## Risk-guided Estimation-aware Acceptance for Learning with Synthetic Data

synthetic proposal $\xrightarrow{\text{target risk } + \text{ normalized score}}$ budgeted acceptance $\xrightarrow{\text{target correction}}$ final estimator.

**Risk-guided**

The design is derived from a scalar surrogate upper bound on target prediction risk

**Acceptance**

An acceptance rule selects candidates subject to a fixed expected budget.

**Estimation-aware**

Candidates are ranked using a normalized metric derived from an initial model trained on real population data only.

**Learning with synthetic data**

Accepted candidates are reweighted so that the synthetic contribution continues to target the real population.

# Target–Proposal Mismatch

Let $Z = (\boldsymbol{X}, Y)$. The real and generated observations generally come from different joint distributions.

### Target Population

$$D_{\text{real}} = \{Z_i\}_{i=1}^{n}, \qquad Z_i \sim p(x, y) = p(y)p(x \mid y).$$

Where prediction performance is evaluated on.

### Synthetic Proposal

$$\widetilde{Z}_j \sim q_{\text{prop}}(x, y) = q(y)q_{\text{gen}}(x \mid y).$$

Induced distribution by proposed label distribution and pre-trained conditional generator.

The joint density ratio measures target relevance of synthetic data:

$$r(x, y) = \frac{p(x, y)}{q_{\text{prop}}(x, y)} = \frac{p(y)}{q(y)} \frac{p(x \mid y)}{q_{\text{gen}}(x \mid y)}.$$

### Main Idea of REAL-Syn

$$\text{acceptance "priority"} \quad \propto \quad \underbrace{r(x, y)}_{\text{target relevance}} \quad \times \quad \underbrace{h(x, y)}_{\text{estimator relevance}}$$

A candidate should be both relevant to the target law and informative for the downstream estimator.

# Acceptance and Target-Corrected Learning

An acceptance rule and its proposal-level budget are

$$\pi(x,y) \in [0,1], \qquad \alpha = \mathbb{E}_{q_{\mathrm{prop}}}\left[\pi(\boldsymbol{X}, Y)\right].$$

The retained synthetic observations follow

$$q_{\mathrm{syn}}\big(x, y; \pi\big) = \frac{q_{\mathrm{prop}}(x,y)\pi(x,y)}{\alpha}.$$

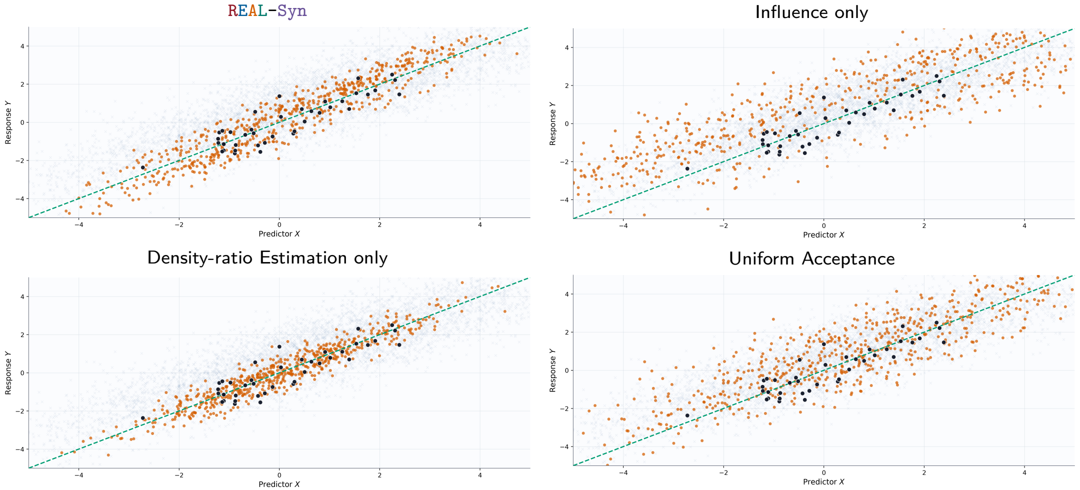
The correction from the accepted distribution back to the target law is

$$W\big(x, y; \pi\big) = \frac{p(x,y)}{q_{\mathrm{syn}}(x,y;\pi)} = \frac{\alpha\, r(x,y)}{\pi(x,y)}.$$

The final estimator minimizes

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} L(Z_i, \boldsymbol{\theta}) + \sum_{j=1}^{\widetilde{M}} \underbrace{W\big(x, y; \pi\big) L(\widetilde{Z}_j, \boldsymbol{\theta})}_{\text{target-corrected synthetic contribution}} \right\}.$$

# A Preview: Example Acceptance Rules




Black: real. Orange: accepted synthetic. Blue: rejected synthetic

# Risk-Guided Starting Point

risk · estimation · acceptance · learning · synthetic proposal

For a prediction model $f_{\boldsymbol{\theta}}$, define the target squared risk

$$R(\boldsymbol{\theta}) = \mathbb{E}_p\left[\{f_{\boldsymbol{\theta}}(\boldsymbol{X}) - Y\}^2\right], \qquad \boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} R(\boldsymbol{\theta}).$$

The loss induces the score and its target covariance:

$$\boldsymbol{\psi}(Z) = \nabla_{\boldsymbol{\theta}} L(Z, \boldsymbol{\theta}^*), \qquad \boldsymbol{V}_p = \mathrm{Var}_p\{\boldsymbol{\psi}(Z)\},$$

$$\boldsymbol{V}_w = \mathrm{Var}_{q_{\mathrm{syn}}}\left\{W(Z;\pi)\boldsymbol{\psi}(Z)\right\}.$$

Under the local M-estimation expansion, with $N = n + \widetilde{M}$, $\rho = \widetilde{M}/N$, and $\mathcal{P}_N$ the augmented law,

$$\mathbb{E}_{\mathcal{P}_N}\left[R(\widehat{\boldsymbol{\theta}})\right] \le \sigma^2 + B^2 + \frac{1}{N}\mathrm{tr}\left[\boldsymbol{A}\left\{(1-\rho)\boldsymbol{V}_p + \rho\boldsymbol{V}_w\right\}\boldsymbol{A}^\top\right] + o(N^{-1}).$$

**Note**

Only $\boldsymbol{V}_w$ depends on the acceptance rule. `REAL-Syn` designs an $\pi$ to control the synthetic contribution to target risk.

# A Scalar Surrogate for Acceptance Rule Design

If $\boldsymbol{V}_p$ is nonsingular, the weighted synthetic covariance satisfies

$$\boldsymbol{V}_w \preceq \eta\big(\pi\big)\boldsymbol{V}_p.$$

The scalar multiplier is

$$\eta\big(\pi\big) = \alpha\, \mathbb{E}_{q_{\mathrm{prop}}}\left[\underbrace{r(\boldsymbol{X}, Y)^2}_{\substack{\text{target}\\ \text{relevance}}} \underbrace{\frac{1}{\pi(\boldsymbol{X}, Y)}}_{\substack{\text{acceptance}\\ \text{``priority''}}} \underbrace{\boldsymbol{\psi}(\boldsymbol{X}, Y)^\top \boldsymbol{V}_p^{-1} \boldsymbol{\psi}(\boldsymbol{X}, Y)}_{\substack{\text{estimation}\\ \text{relevance}}}\right].$$

Consequently,

$$\mathbb{E}_{\mathcal{P}_N}\left[R(\widehat{\boldsymbol{\theta}})\right] \le \sigma^2 + B^2 + \frac{1}{N}\left\{1 - \rho + \rho\,\eta\big(\pi\big)\right\} \operatorname{tr}(\boldsymbol{A}\boldsymbol{V}_p\boldsymbol{A}^\top) + o(N^{-1}).$$

## Fixed-budget functional optimization

$$\min_{\pi}\ \eta\big(\pi\big) \quad \text{subject to} \quad \mathbb{E}_{q_{\mathrm{prop}}}\left[\pi(\boldsymbol{X}, Y)\right] = \alpha, \qquad 0 \le \pi(x, y) \le 1.$$

# The REAL-Syn Acceptance Rule

**R**isk-guided **E**stimation-aware **A**cceptance for **L**earning with **Syn**thetic Data

The risk-guided surrogate optimization gives

$$\boxed{\pi^\star(x,y) = \min\left\{1,\ c\,\underbrace{r(x,y)}_{\substack{\text{target}\\\text{relevance}}}\,\underbrace{\sqrt{\boldsymbol{\psi}(x,y)^\top \boldsymbol{V}_p^{-1}\boldsymbol{\psi}(x,y)}}_{\substack{\text{estimator}\\\text{relevance}}}\right\}}$$

where $c > 0$ is calibrated so that

$$\mathbb{E}_{q_{\mathrm{prop}}}\left[\pi^\star(\boldsymbol{X}, Y)\right] = \alpha.$$

### Risk-guided

Target-risk bound determines the optimization criterion.

### Acceptance

Truncating at one respects proposal availability and makes $\pi^\star$ a valid probability.

### Corrected learning

Accepted candidates receive

$$W^\star(z) = \frac{\alpha r(z)}{\pi^\star(z)}.$$

# Intuition: Relationship of Estimator Relevance with Cook's Distance

For squared-loss linear regression $L(\boldsymbol{x}, y), \boldsymbol{\theta}) = (y - \boldsymbol{x}^\top \boldsymbol{\theta})^2$, define the pilot residual and leverage

$$e(\boldsymbol{x}, y) = y - \boldsymbol{x}^\top \widehat{\boldsymbol{\theta}}, \qquad h(\boldsymbol{x}) = \boldsymbol{x}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x},$$

where $d_\beta$ is the number of fitted coefficients and $\widehat{\sigma}^2$ is the residual-variance estimate.

## Classical Cook's distance

For an in-sample observation,

$$\widehat{D}_{\mathrm{Cook}}(\boldsymbol{x}, y) = \frac{e(\boldsymbol{x}, y)^2}{d_\beta \widehat{\sigma}^2} \frac{h(\boldsymbol{x})}{\{1 - h(\boldsymbol{x})\}^2}.$$

Measures the effect of deleting an observation from the fitted regression.

## REAL-Syn estimator relevance metric

The normalized score is evaluated as

$$\widehat{D}_{\mathrm{RL}}(\boldsymbol{x}, y) = \widehat{\boldsymbol{\psi}}(\boldsymbol{x}, y)^\top \widehat{\boldsymbol{V}_p}^{-1} \widehat{\boldsymbol{\psi}}(\boldsymbol{x}, y) = n \frac{e(\boldsymbol{x}, y)^2}{\widehat{\sigma}^2} h(\boldsymbol{x}).$$

Measures a synthetic candidate's standardized contribution to coefficient estimation.

Therefore,

$$\widehat{D}_{\mathrm{RL}} = n d_\beta \{1 - h(\boldsymbol{x})\}^2 \widehat{D}_{\mathrm{Cook}} \implies \boxed{\texttt{REAL-Syn}\text{ acceptance} \propto \text{target relevance} \cdot \sqrt{\text{Cook's-type Metric}}}$$

# How REAL-Syn is Implemented on Real World Applications

1. **Estimate target relevance.** Estimate the joint density ratio $\widehat{r}(x,y) \approx p(x,y)/q_{\text{prop}}(x,y)$,
2. **Fit an initial model on real data.**. Compute the plug-in scores $\widehat{\boldsymbol{\theta}}_{\text{init}}$ and $\widehat{\psi}(x,y) = \nabla_{\boldsymbol{\theta}} L((x,y), \widehat{\boldsymbol{\theta}}_{\text{init}})$
3. **Estimate the target normalized score.** Compute the empirical covariance of the real-data scores:
$$\widehat{\boldsymbol{V}}_p = \widehat{\operatorname{Var}}_p \left\{ \widehat{\psi}(Z) \right\}.$$
4. **"Rank" each synthetic candidate.** For every synthetic candidate $(x,y)$ calculate
$$\widehat{s}(x,y) = \widehat{r}(x,y) \left\{ \widehat{\psi}(x,y)^{\top} \widehat{\boldsymbol{V}}_p^{-1} \widehat{\psi}(x,y) \right\}^{1/2}$$
5. **Calibrate and apply the acceptance rule.** Set $\widehat{\pi}(x,y) = \min\left\{1,\, c\,\widehat{s}(x,y)\right\}$, where $c > 0$ is chosen so that $\frac{1}{M}\sum_{j=1}^{M} \widehat{\pi}(\widetilde{Z}_j) = \alpha$.. Accept candidate $j$ independently with probability $\widehat{\pi}(\widetilde{Z}_j)$.
6. **Correct and refit the final model.** Assign each accepted candidate the weight
$$\widehat{W}(x,y) = \frac{\alpha\,\widehat{r}(x,y)}{\widehat{\pi}(x,y)},$$
and combine the weighted accepted candidates with the real observations to fit $\widehat{\boldsymbol{\theta}}_{\text{final}}$.

Section 3

# Numerical Simulations

# Misspecified Linear Regression Simulation Design

The simulations isolate sample-level acceptance by using an oracle density ratio.

**Target distribution**

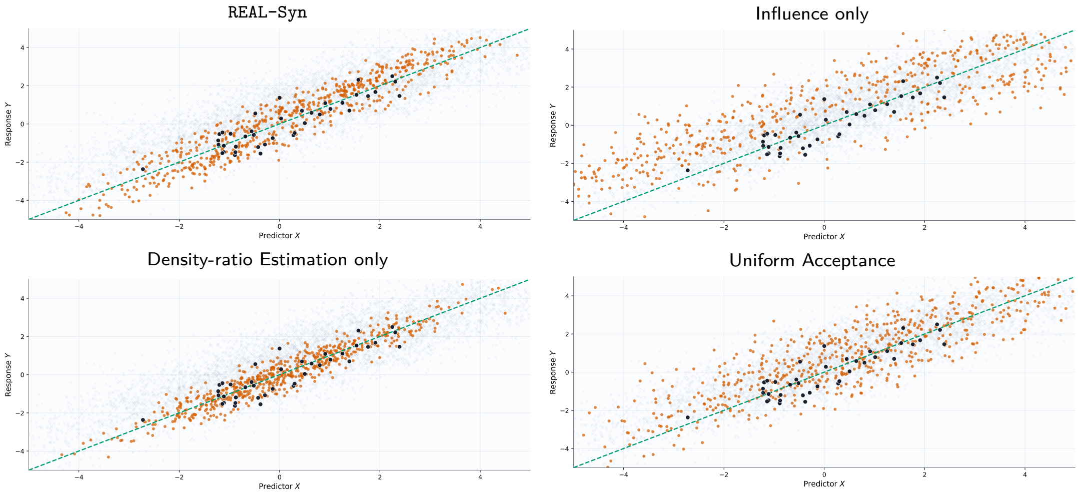
$$\boldsymbol{X} \sim N(0, 1.5^2 I_d), \qquad Y = \boldsymbol{X}^\top \boldsymbol{\beta} + \varepsilon,$$
$$\varepsilon \sim N(0, 0.5^2), \quad d \in \{1, 2, 5\}.$$

**Synthetic proposal**

$$\widetilde{\boldsymbol{X}} \sim N(0.5\mathbf{1}_d, 2.5^2 I_d),$$
$$\widetilde{Y} = 0.5 + \widetilde{\boldsymbol{X}}^\top (0.8\boldsymbol{\beta}) + \widetilde{\varepsilon}, \qquad \widetilde{\varepsilon} \sim N(0, 1^2).$$

- $n = 40$ real training samples; 6000 candidate synthetic samples.
- Acceptance budget $\alpha = 0.10$; 100 repetitions.
- Baselines: real only, uniform acceptance, density-ratio only, influence only.

# Example of Acceptance Patterns in Misspecified Linear Regression $(d = 1)$




Black: real. Orange: accepted synthetic. Blue: rejected synthetic

# Simulation Results of Misspecified Linear Regression $(d = 1)$

Table: $d = 1$ Population MSE and Coefficient covariance trace

| Method | Pop. MSE$\pm$*MCSE* | $\mathrm{tr}\left\{\mathrm{Cov}(\widehat{\beta})\right\} \times 10^3$ |
|---|---|---|
| Real-only | $0.2618 \pm 1.44\times10^{-3}$ | $7.813$ |
| Uniform | $0.2510 \pm 1.01\times10^{-4}$ | $0.737$ |
| Density-ratio only | $0.2508 \pm 8.00\times10^{-5}$ | $0.564$ |
| Influence only | $0.2535 \pm 3.67\times10^{-4}$ | $2.643$ |
| **REAL-Syn** | $\mathbf{0.2505 \pm 5.50\times10^{-5}}$ | $\mathbf{0.365}$ |

# Simulation Results of Strongest Baseline (DRE-only) vs REAL-Syn

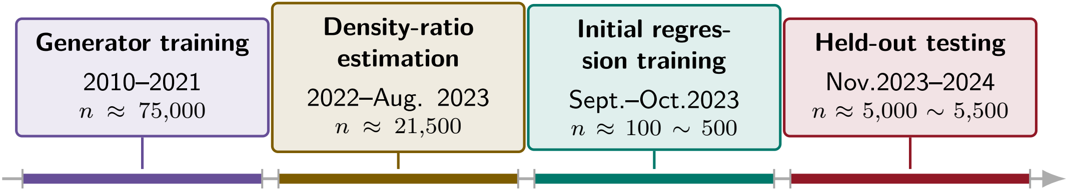
| $d$ | $\mathrm{tr}\left\{\mathrm{Cov}(\widehat{\boldsymbol{\beta}})\right\} \times 10^3$ | Inference $\times 10^5$ | Pop. Test MSE |
|---|---|---|---|
| 1 | DRE-only $0.564 \rightarrow$ REAL-Syn $0.365$<br>$-35\%$ | $8.0 \rightarrow 5.5$<br>$-31\%$ | $0.2508 \rightarrow 0.2505$<br>$-0.12\%$ |
| 2 | DRE-only $0.747 \rightarrow$ REAL-Syn $0.578$<br>$-23\%$ | $10.3 \rightarrow 7.2$<br>$-30\%$ | $0.2512 \rightarrow 0.2508$<br>$-0.16\%$ |
| 5 | DRE-only $1.170 \rightarrow$ REAL-Syn $1.030$<br>$-12\%$ | $15.0 \rightarrow 12.1$<br>$-19\%$ | $0.2521 \rightarrow 0.2519$<br>$-0.08\%$ |

## Section 4

# Solar Weather Forecasting Application

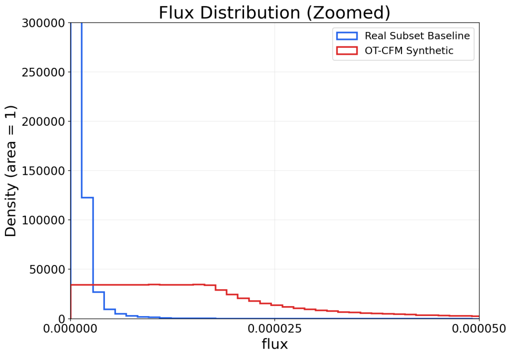
# Solar Flare Regression Setting

- Goal is to predict standardized log soft X-ray flux $y$ from 20 dimensional SHARPs $\boldsymbol{x}$.

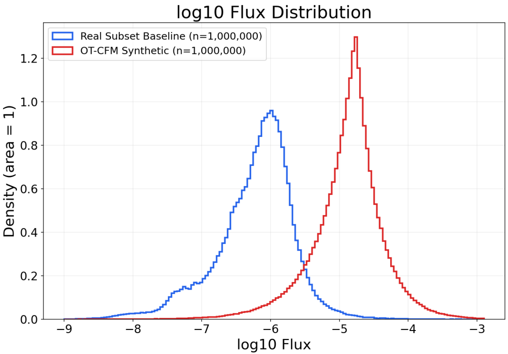



- Conditional flow-matching generator $q_{\text{gen}}(\boldsymbol{x} \mid y)$ produces SHARP vectors conditioned on flux.
- All post-2021 data are standardized using generator-training statistics.
- Density ratio $r(x, y)$ is estimated by LightGBM source discrimination with Platt calibration.
- Each combination (**60** total) of fixed chronological train test split with a varying $M$ large synthetic pool is replicated 150 times. In a given replicate, all methods share the same synthetic pool.
- Variation among replicates come from DRE training splits, synthetic generation, and randomized acceptance. All experiments were run with budget $\alpha = \mathbb{E}_{q_{\text{prop}}}[\pi(\boldsymbol{X}, Y)] = 0.05$.

# Solar Synthetic Proposal Marginal



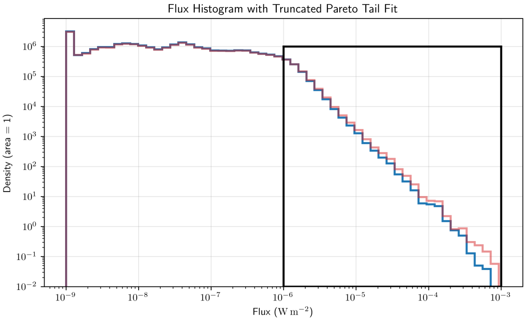

Raw flux scale



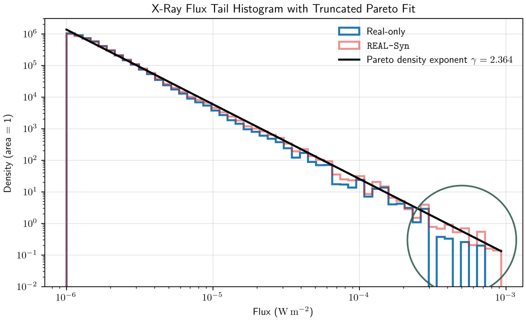

Log10 flux scale

The proposal response distribution intentionally over-represents rare high-flux events before conditional SHARP generation.

# Synthetic Augmentation Increases Rare Event Coverage in the Tail Region




Observed and pooled flux distribution




Tail fit and sparse-tail interpolation

# Counts of Method Wins and Test MSE $\pm 1\hat{\sigma}$ Results

A method is considered to "win" if it achieves the lowest test MSE without overlapping $\pm 1\widehat{\sigma}$ error bounds. If overlaps occur, the smallest coefficient covariance serves as the tiebreaker.

| Method | Uniform | DRE-only | Influence-only | `REAL-Syn` |
|---|---|---|---|---|
| Wins (out of **60**) | $3$ | $19$ | $2$ | $\mathbf{36}$ |

# Counts of Method Wins and Test MSE $\pm 1\hat{\sigma}$ Results

A method is considered to "win" if it achieves the lowest test MSE without overlapping $\pm 1\widehat{\sigma}$ error bounds. If overlaps occur, the smallest coefficient covariance serves as the tiebreaker.

| Method | Uniform | DRE-only | Influence-only | `REAL-Syn` |
|---|---|---|---|---|
| Wins (out of **60**) | 3 | 19 | 2 | **36** |

Fixed downstream training size $n \approx 250$. Entries are held-out standardized log-flux test MSE, mean $\pm$ one standard deviation across 150 replicates.

| | Generated synthetic candidate-pool size | | | | | |
|---|---|---|---|---|---|---|
| Method | $M \approx 5$ | $M \approx 25$ | $M \approx 50$ | $M \approx 125$ | $M \approx 250$ | $M \approx 500$ |
| Real-only | $8.298 \pm 0.000$ | $8.298 \pm 0.000$ | $8.298 \pm 0.000$ | $8.298 \pm 0.000$ | $8.298 \pm 0.000$ | $8.298 \pm 0.000$ |
| Uniform | $8.185 \pm 0.638$ | $7.718 \pm 1.510$ | $7.344 \pm 1.964$ | $5.254 \pm 2.909$ | $3.348 \pm 2.866$ | $2.001 \pm 2.210$ |
| DRE-only | $8.095 \pm 0.974$ | $5.008 \pm 2.830$ | $2.478 \pm 2.404$ | $\mathbf{0.732} \pm 0.913$ | $0.353 \pm 0.240$ | $0.284 \pm 0.028$ |
| Influence-only | $8.138 \pm 0.973$ | $7.803 \pm 1.322$ | $6.931 \pm 2.186$ | $5.725 \pm 2.711$ | $3.998 \pm 2.861$ | $2.073 \pm 2.118$ |
| `REAL-Syn` | $\mathbf{8.029} \pm 1.169$ | $\mathbf{4.975} \pm 2.739$ | $\mathbf{2.447} \pm 2.260$ | $0.736 \pm 0.970$ | $\mathbf{0.349} \pm 0.221$ | $\mathbf{0.283} \pm 0.026$ |

# Empirical Coefficient Variability

Trace of empirical covariance of fitted coefficients across 150 replicates, intercept included. Because the chronological real training set is fixed, real-only has conditional covariance zero and is omitted below.

| | Generated synthetic candidate-pool size | | | | | |
|---|---|---|---|---|---|---|
| Method | 5 | 25 | 50 | 125 | 250 | 500 |
| Uniform | **135.62** | 792.88 | **1303.90** | 2859.30 | 2971.70 | 2116.90 |
| DRE-only | 348.79 | 3437.50 | 2512.30 | **699.64** | 205.62 | 47.62 |
| Influence-only | 332.18 | **425.01** | 1660.60 | 2731.40 | 2868.60 | 1816.60 |
| **REAL-Syn** | 506.51 | 3234.70 | 2445.20 | 713.58 | **201.32** | **46.25** |

Section 5

# Discussion and Next Steps

# Main Takeaways

1. As our target remains the real-data population, synthetic samples are only useful in the sense of improving the target estimator after correcting for any proposal mismatch.
2. Importance weighting alone gives a correction identity and should not be only criterion to decide which synthetic candidates should enter training.
3. Our derived acceptance for REAL-Syn rule has two components:

$$\text{acceptance “priority”} \quad \propto \quad \underbrace{r(x,y)}_{\text{target relevance}} \times \underbrace{\{\psi(x,y)^\top \boldsymbol{V}_p^{-1}\psi(x,y)\}^{1/2}}_{\text{estimator relevance: “Cook-type” metric}} .$$

4. Simulations of misspecified regression case show that REAL-Syn outperforms various other methods and ablation baselines.
5. For X-Ray flux on SHARPs regresssion, “correction-aware” selections (REAL-Syn and DRE-only) dominate uniform and influence-only augmentation, with REAL-Syn being more consistent and stable

# Limitations and Potential Next Steps

- Current theoretical results are an asymptotic surrogate-risk analysis and not an exact finite-sample oracle inequality.
- Extend `REAL-Syn` to high-dimensional and dependent-data regimes, and develop task-aware selection rules for nonlinear models like neural networks.
- How to leverage the known "cyclical" solar cycle information and treat solar weather forecasting as a label shift problem $p_t(y)$; currently working on how to use a "well-trained" conditional generator $p_{\text{gen}}(x \mid y)$ along with this known external information to improve performance

# Acknowledgments


- I would to formally thank our collaborators Zijie Tian and Thomas Lee from UC Davis and my advisor Yang Chen.
- This work has been supported by the NASA Space Weather Center of Excellence program at the University of Michigan under award 80NSSC23M0191 (CLEAR Center).
- This research was also supported in part through computational resources and services provided by Advanced Research Computing (ARC) at the University of Michigan.

Thank you!

Section 6

# Appendix

# Assumptions and Regularity Conditions

- Real samples are i.i.d. from $p$; accepted synthetic samples are conditionally i.i.d. from $q_{\mathrm{syn}}(\cdot;\pi)$ for fixed $\widetilde{M}$.
- The M-estimator admits a local linear expansion at $\boldsymbol{\theta}^*$, with controlled second-order remainder.
- Local smoothness of the predictor via twice continuously differentiable in a neighborhood of $\theta^*$.

# Proof sketch for the clipped square-root rule

Write

$$a(z) = r(z)^2\, \psi(z)^\top \boldsymbol{V}_p^{-1} \psi(z), \qquad z = (x, y).$$

The surrogate problem is equivalent to minimizing

$$\mathbb{E}_q \left[ \frac{a(Z)}{\pi(Z)} \right] \quad \text{subject to} \quad \mathbb{E}_q \pi(Z) = \alpha, \quad 0 \le \pi \le 1.$$

Ignoring the upper constraint, the pointwise first-order condition gives

$$-\frac{a(z)}{\pi(z)^2} + \lambda = 0 \quad \Longrightarrow \quad \pi(z) = \sqrt{a(z)/\lambda}.$$

Imposing $\pi \le 1$ clips the solution:

$$\pi^\star(z) = \min\{1, c\sqrt{a(z)}\} = \min\left\{1, c\, r(z) \{\psi(z)^\top \boldsymbol{V}_p^{-1} \psi(z)\}^{1/2} \right\}.$$

The scalar $c$ is chosen to satisfy the budget $\mathbb{E}_q \pi^\star = \alpha$.

# SHARPs

| Parameter | Description |
| --- | --- |
| USFLUX | Total unsigned radial magnetic flux |
| MEANGAM | Mean inclination angle of the field from the local radial direction |
| MEANGBT | Mean horizontal gradient of the total magnetic field strength |
| MEANGBZ | Mean horizontal gradient of the radial/vertical field component |
| MEANGBH | Mean horizontal gradient of the horizontal field component |
| MEANJZD | Mean vertical current density |
| TOTUSJZ | Total unsigned vertical current |
| MEANALP | Mean characteristic twist parameter |
| MEANJZH | Mean vertical-current-helicity proxy |
| TOTUSJH | Total unsigned vertical-current-helicity proxy |
| ABSNJZH | Absolute value of the net vertical-current-helicity proxy |
| SAVNCPP | Sum of absolute net vertical currents over positive and negative polarities |
| MEANPOT | Mean photospheric excess/free magnetic energy density |
| TOTPOT | Surface integral/proxy of photospheric excess/free magnetic energy density |
| MEANSHR | Mean shear angle between observed and potential magnetic field |
| SHRGT45 | Percentage of area/pixels with shear angle greater than $45^\circ$ |
| SIZE | Projected area of the patch on the image |
| SIZE_ACR | Projected area of active pixels on the image |
| NACR | Number of active pixels in the patch |
| NPIX | Number of pixels within the patch |

# Logistic Regression Example

Extending to higher dimensional OLS: Using $d \in \{10, 20\}$ with $n_{\text{real}} = 200$, $M = 30000$ and $\alpha = 0.1$, and regimes of 1) sparse signal padded with zeros and 2) dense norm-matched signal $\beta = \sqrt{2.2/d}\mathbf{1}_d$.

| Setting | DRE-only Test MSE (MC SE) | REAL-Syn Test MSE (MC SE) | DRE-only Cov. $\times 10^3$ | REAL-Syn Cov. $\times 10^3$ |
|---|---|---|---|---|
| Sparse, $d = 10$ | $0.24544\,(6.45\times10^{-5})$ | $0.24543\,(6.59\times10^{-5})$ | 0.666 | 0.669 |
| Sparse, $d = 20$ | $0.27150\,(8.86\times10^{-4})$ | $0.27151\,(8.83\times10^{-4})$ | 10.911 | 10.913 |
| Dense, $d = 10$ | $0.24544\,(6.18\times10^{-5})$ | $0.24545\,(6.43\times10^{-5})$ | 0.701 | 0.703 |
| Dense, $d = 20$ | $0.27075\,(8.57\times10^{-4})$ | $0.27076\,(8.53\times10^{-4})$ | 10.524 | 10.531 |

# Equal-budget fixed-pool baselines

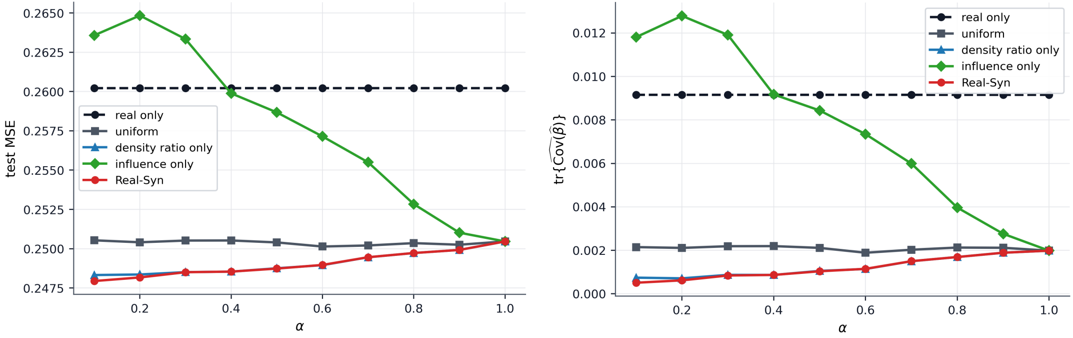
Set up using $d = 5$ setting, $n = 40$, one shared pool of $M = 6000$ candidates, fitted synthetic budget $B = 600$, and 100 repetitions.

| Method | Unique | ESS | Test MSE | Cov. $\times 10^3$ |
|---|---|---|---|---|
| DRE-only | 600 | 522 | 0.248444 | 1.170 |
| REAL-Syn | 598 | 222 | **0.248191** | **1.030** |
| Multinomial SIR | 432 | 326 | 0.249387 | 1.802 |
| Residual SIR | 462 | 377 | 0.249548 | 1.792 |
| Stratified SIR | 507 | 438 | 0.249219 | 1.698 |
| Subsampled SNIS | 600 | 72 | 0.254031 | 4.571 |
| Clipped SNIS | 600 | 75 | 0.253960 | 4.534 |

Multinomial SIR uses classical sampling-importance-resampling (Rubin, 1987), Residual and Stratified SIR use standard variance-reduced resampling (Douc et al., 2005), Subsampled SNIS uniformly selects $600$ unique candidates and self-normalizes their ratios, and Clipped SNIS uses the same subset with threshold $\bar{r}_S\sqrt{600}$ (Ionides, 2008).

# Acceptance-budget sensitivity: full grid

Expected accepted count is fixed at 500; changing $\alpha$ changes selectivity and candidate-pool size.



# Controlled density-ratio misspecification

Submitted $d = 5$ setting; only the ratio used for acceptance and correction is changed to $r_\lambda = (1 - \lambda)r + \lambda$. Entries are Test MSE and coefficient covariance trace $\times 10^3$.

| $\lambda$ | DRE-only MSE | REAL-Syn MSE | DRE-only Cov. | REAL-Syn Cov. |
|---|---|---|---|---|
| 0 | 0.2484 | **0.2482** | 1.170 | **1.030** |
| 0.05 | 0.2509 | **0.2505** | 1.663 | **1.309** |
| 0.10 | 0.2572 | **0.2563** | 2.088 | **1.444** |
| 0.15 | **0.2646** | 0.2651 | 2.270 | **1.784** |
| 0.20 | **0.2743** | 0.2756 | 2.514 | **2.070** |

# Binary Class Gaussian Logistic Regression Simualtion

5 dimensional class-conditional Gaussian classification, exact joint ratio, $n = 100$, $M = 6000$, $\alpha = 0.1$, 100 repetitions.

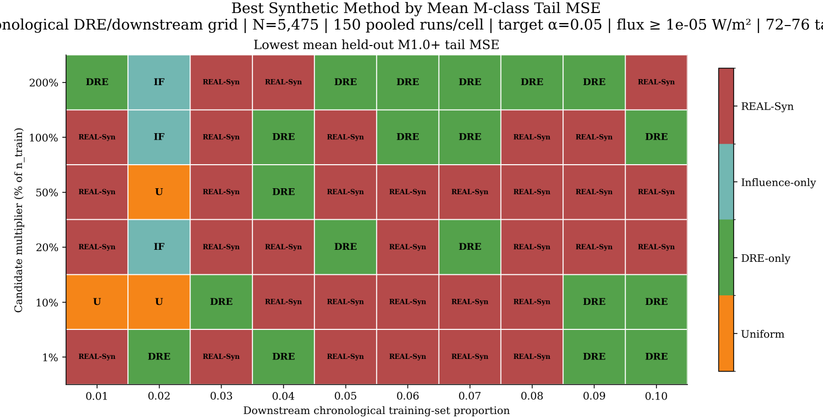
| $(\pi_p, \pi_q)$ | DRE-only Test log-loss (MC SE) | REAL-Syn Test log-loss (MC SE) | DRE-only Cov. $\times 10^3$ | REAL-Syn Cov. $\times 10^3$ |
|---|---|---|---|---|
| $(0.2, 0.5)$ | $0.3795\,(2.39\times10^{-4})$ | $\mathbf{0.3786}\,(\mathbf{1.68}\times\mathbf{10^{-4}})$ | $36.189$ | $\mathbf{28.737}$ |
| $(0.5, 0.5)$ | $0.5111\,(2.48\times10^{-4})$ | $\mathbf{0.5107}\,(\mathbf{2.23}\times\mathbf{10^{-4}})$ | $25.420$ | $\mathbf{23.787}$ |

# Fixing sythentic pool size, changing training size for solar experiments

Fixed synthetic candidate pool size to be half the size of the training samples.

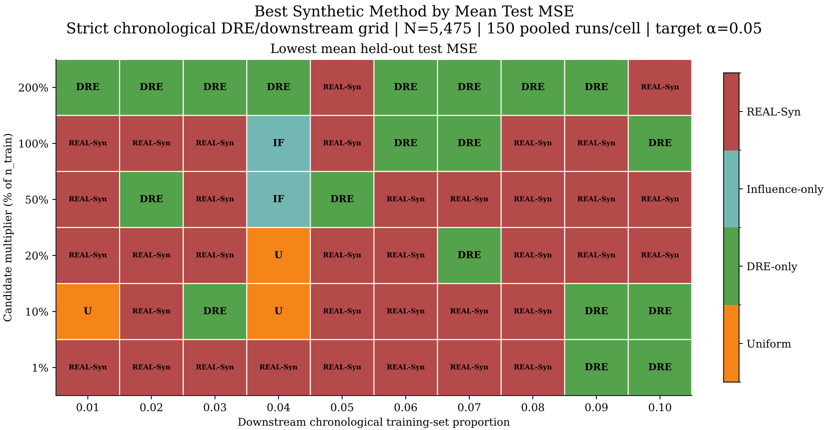
| | Real Training-Set Size ($n$) | | | | |
|---|---|---|---|---|---|
| Method | $n \approx 50$ | $n \approx 150$ | $n \approx 250$ | $n \approx 350$ | $n \approx 450$ |
| Real-only | $84.382 \pm 0.000$ | $2.548 \pm 0.000$ | $8.298 \pm 0.000$ | $15.234 \pm 0.000$ | $3.892 \pm 0.000$ |
| Uniform | $82.784 \pm 20.999$ | $2.017 \pm 0.759$ | $5.254 \pm 2.909$ | $10.076 \pm 5.314$ | $2.919 \pm 1.140$ |
| DRE-only | $68.671 \pm 44.810$ | $0.615 \pm 0.572$ | $\mathbf{0.732} \pm 0.913$ | $1.425 \pm 1.893$ | $\mathbf{0.733} \pm 0.707$ |
| Influence-only | $84.204 \pm 15.623$ | $2.128 \pm 0.693$ | $5.725 \pm 2.711$ | $9.187 \pm 5.084$ | $2.879 \pm 1.124$ |
| REAL-Syn | $\mathbf{66.129} \pm 44.400$ | $\mathbf{0.595} \pm 0.496$ | $0.736 \pm 0.970$ | $\mathbf{1.416} \pm 1.921$ | $\mathbf{0.733} \pm 0.709$ |

# M-class tail Test MSE Heatmap




Tail MSE is computed on test observations above the M-class threshold $10^{-5}\ \mathrm{W/m^2}$ (about 75 tail observations per cell).

# Best Performing Method for a Combo via global test MSE




Each cell reports the synthetic method with the lowest mean held-out test MSE under the strict chronological DRE/downstream grid. REAL-Syn and DRE-only dominates.

# Counts of Best method M-class Tail MSE

| Method | Uniform | DRE-only | Influence-only | **REAL-Syn** |
|---|---|---|---|---|
| Wins (out of 60) | 3 | 20 | 3 | **34** |

# Chronological solar result: actual M-class tail-MSE row

Fixed downstream training size $n \approx 250$. Entries are M-class tail MSE, mean $\pm$ one standard deviation across 150 paired replicates.

| | Generated synthetic candidate-pool size | | | | | |
|---|---|---|---|---|---|---|
| Method | $M \approx 5$ | $M \approx 25$ | $M \approx 50$ | $M \approx 125$ | $M \approx 250$ | $M \approx 500$ |
| Real-only | $26.064 \pm 0.000$ | $26.064 \pm 0.000$ | $26.064 \pm 0.000$ | $26.064 \pm 0.000$ | $26.064 \pm 0.000$ | $26.064 \pm 0.000$ |
| Uniform | $25.817 \pm 1.410$ | $24.733 \pm 3.669$ | $23.851 \pm 4.788$ | $18.748 \pm 7.518$ | $13.795 \pm 7.749$ | $9.887 \pm 6.606$ |
| DRE-only | $25.593 \pm 2.366$ | $18.148 \pm 7.444$ | $\mathbf{11.371} \pm 6.974$ | $5.699 \pm 3.566$ | $3.859 \pm 1.538$ | $\mathbf{3.092} \pm 0.428$ |
| Influence-only | $25.680 \pm 2.403$ | $24.962 \pm 3.051$ | $22.896 \pm 5.376$ | $19.935 \pm 7.010$ | $15.545 \pm 7.692$ | $10.271 \pm 6.333$ |
| REAL-Syn | $\mathbf{25.428} \pm 2.891$ | $\mathbf{18.147} \pm 7.149$ | $11.400 \pm 6.616$ | $\mathbf{5.698} \pm 3.663$ | $\mathbf{3.834} \pm 1.488$ | $3.095 \pm 0.426$ |