

# DREA: A Probabilistic Distributional Attribution Framework for Trustworthy Explanations in Diffusion-Based Predictive Models

Chibuike Chiedozie Ibebuchi

`chibuike.ibebuchi@morgan.edu`

Morgan State University <https://orcid.org/0000-0001-6010-2330>

Qunying Huang

University of Wisconsin-Madison <https://orcid.org/0000-0003-3499-7294>

---

## Research Article

**Keywords:** diffusion models, generative forecasting, uncertainty quantification, explainable AI, distributional attribution, SHAP

**Posted Date:** June 11th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-9969649/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** The authors declare no competing interests.

---

# **DREA: A Probabilistic Distributional Attribution Framework for Trustworthy Explanations in Diffusion-Based Predictive Models**

Chibuike Chiedozie Ibebuchi<sup>1,2\*</sup>, Qunying Huang<sup>3</sup>

<sup>1</sup>Department of Mathematics, Morgan State University, Baltimore, Maryland, United States

<sup>2</sup>Center for Urban and Coastal Climate Science Research, Morgan State University, United States

<sup>3</sup>Department of Geography, University of Wisconsin-Madison

Corresponding author: Chibuike Chiedozie Ibebuchi, (Chibuike.ibebuchi@morgan.edu)  
Corresponding author orcid: <https://orcid.org/0000-0001-6010-2330>

## **Abstract**

Diffusion-based generative models can produce predictive distributions, but their explanations are often derived from point-attribution methods that do not quantify how predictors influence uncertainty, distributional shape, or extreme-event behavior. We propose Distributional Reliability Explanation Attribution (DREA), a probabilistic attribution framework for diffusion-based predictive models. Instead of assigning a single importance score, DREA estimates how reliably predictor perturbations alter properties of generated predictive distributions across repeated bootstrap realizations. The framework decomposes attribution into distributional shift, uncertainty width, variance, tail-risk behavior, and probabilistic skill using empirical reliability frequencies. We validate DREA with a controlled synthetic benchmark in which separate predictors control the conditional mean, predictive variance, upper-tail behavior, correlated proxy information, and irrelevant noise. DREA recovers the known distributional roles: mean drivers are identified through distributional-shift and skill components, uncertainty drivers through width and variance components, and tail drivers through tail-risk components. This decomposition shows not only which predictors matter, but which aspect of the predictive distribution they control. DREA also assigns near-zero reliability to irrelevant noise and lower reliability to the Mean-proxy variable than to the true Mean driver. We further demonstrate DREA on diffusion-based hydrological forecasting using streamflow, meteorological, and basin predictors. The real-world experiment shows that DREA decomposes predictor influence into interpretable probabilistic components associated with predictive skill, uncertainty amplification, variance control, and upper-tail discharge risk. Event-level analyses provide local explanations by estimating how often predictor perturbations change predicted discharge and flood-tail probability. DREA extends explainability from point-wise feature importance to probabilistic distributional attribution for uncertainty-sensitive decision support.

**Keywords:** diffusion models; generative forecasting; uncertainty quantification; explainable AI; distributional attribution; SHAP

## Highlights

- Introduced DREA for explaining diffusion-based predictive distributions.
- Decomposed attribution into shift, width, variance, tail-risk, and skill components.
- Recovered known mean, uncertainty, and tail drivers in a synthetic benchmark.
- Distinguished true distributional drivers from correlated proxy variables.
- Demonstrated global and event-level explanations for hydrological forecasting.

## 1. Introduction

Probabilistic forecasting has become increasingly important in environmental prediction, climate science, hydrology, energy systems, and other uncertainty-sensitive applications where decisions depend not only on expected outcomes, but also on predictive uncertainty and extreme-event risk [1-4]. Recent advances in diffusion models and related generative forecasting systems have enabled the production of realistic predictive distributions that capture multimodality, stochastic variability, and tail behavior more effectively than many deterministic forecasting approaches [4,5]. These models are therefore increasingly being applied to high-impact forecasting and decision-support tasks, including weather forecasting [4] and beyond [5].

Despite these advances, explainability methods for generative AI remain less mature than those for traditional predictive models and often inherit assumptions from deterministic or discriminative machine-learning paradigms [6]. Widely used approaches such as Shapley Additive Explanations (SHAP) [7], permutation importance [8], integrated gradients [9], and related local attribution methods [10] primarily estimate how predictors contribute to changes in point predictions or averaged model outputs. While these approaches have proven useful for conventional regression and classification tasks, they primarily attribute model outputs or individual predictions to input features and do not explicitly quantify how predictors influence separate properties of predictive distributions, such as uncertainty width, variance amplification, distributional shift, or extreme-tail behavior [7-11].

This limitation becomes particularly important in diffusion-based forecasting systems because the central output of the model is not a single prediction, but an ensemble of

generated predictive realizations [12,13]. In such settings, two predictors may produce similar mean attribution scores while affecting fundamentally different aspects of the predictive distribution. For example, one predictor may primarily shift the median prediction, another may amplify uncertainty variance, while a third may specifically alter upper-tail risk. Conventional point-attribution approaches often collapse these distinct probabilistic behaviors into a single importance value, potentially obscuring the mechanisms through which predictors influence uncertainty-sensitive forecasts [7-11].

An additional challenge in explainable forecasting is the presence of correlated predictors and proxy variables. Attribution methods based on predictive association can assign high importance to variables that are strongly correlated with the true driver, even when they do not directly control the underlying probabilistic behavior of the forecasting system [14-16]. This problem becomes especially critical in environmental forecasting, where meteorological, hydrological, and spatial predictors frequently exhibit strong multicollinearity and temporal dependence [14].

To address these limitations, we propose Distributional Reliability Explanation Attribution (DREA), a probabilistic explainability framework for diffusion-based and generative forecasting systems. Rather than assigning a single importance score to each predictor, DREA quantifies how reliably perturbations to predictors alter different properties of generated predictive distributions across repeated perturbation realizations. The framework decomposes explanation into multiple probabilistic components. By evaluating empirical reliability frequencies across repeated generated samples and perturbation replicates, DREA transforms attribution from deterministic importance estimation into a probabilistic measure of distributional control.

Evaluating explainability methods is challenging because real-world datasets rarely provide known ground-truth attribution structures. Consequently, it is often impossible to determine whether an explanation method correctly identifies the true mechanisms governing predictive behavior [17-20]. In response, we first validate DREA using a controlled synthetic benchmark in which the underlying data-generating processes are explicitly known. This controlled design allows direct evaluation of whether DREA correctly recovers known probabilistic roles and whether it distinguishes true distributional drivers from correlated proxies and irrelevant predictors.

We further demonstrate the framework on a diffusion-based hydrological forecasting application that integrates meteorological and basin predictors to generate probabilistic streamflow forecasts at United States Geological Survey (USGS) stream gaging stations across Maryland [21]. The hydrology experiment serves as a real-world demonstration of DREA in decomposing predictor influence into interpretable probabilistic components associated with predictive skill, uncertainty amplification, variance control, and upper-tail discharge

risk. Beyond global explanations, the framework additionally enables event-level probabilistic attribution by estimating how often predictor perturbations increase or decrease predicted discharge and flood-tail probabilities across generated diffusion samples.

The main contributions of this study are therefore fourfold:

1. We introduce DREA, a probabilistic attribution framework for generative forecasting systems that explains predictor influence on predictive distributions rather than only point predictions.
2. We propose a generalized multi-metric decomposition framework that separates predictor influence into distributional shift, uncertainty width, variance, tail-risk behavior, and probabilistic skill components.
3. We develop a controlled synthetic benchmark with known probabilistic generating mechanisms to rigorously evaluate distributional explainability and compare DREA against conventional attribution approaches.
4. We demonstrate the applicability of DREA in a real-world diffusion-based hydrological forecasting system, including both global probabilistic explanations and event-level attribution analyses.

Overall, this work extends explainability from deterministic feature importance toward probabilistic distributional attribution, providing a more reliable framework for interpreting generative forecasting systems in uncertainty-sensitive applications. By converting distributional sensitivity into interpretable reliability probabilities, DREA can serve as an explanation module for expert and intelligent decision-support systems where uncertainty, risk, and extreme-event behavior influence operational decisions.

## **2. Data and Methods**

DREA consists of six major components (Figure 1): (1) study design and predictive datasets, where the framework is evaluated using both a controlled synthetic benchmark with known predictor roles and a real-world diffusion-based hydrological forecasting application to assess its ability to provide probabilistic distributional attribution for uncertainty, predictive skill, and extreme-event behavior; (2) model training and prediction within a conditional diffusion framework that produces full predictive distributions, enabling probabilistic rather than point-wise forecasting, and benchmark the approach against standard baseline models; (3) global and event-level predictor perturbation, which evaluates how predictor changes affect generated distributions; (4) distributional comparison using metrics such as Wasserstein distance, predictive interval width, variance, tail probability, and CRPS to quantify distributional sensitivity; (5) probabilistic attribution estimation, which measures the reliability that a predictor induces material distributional changes across repeated

bootstrap realizations; and (6) outputs and interpretation, including global predictor rankings and local event-level explanations of uncertainty, predictive skill, and extreme-event risk. Together, these components extend explainability from point-wise feature importance to uncertainty-aware probabilistic attribution for diffusion-based predictive modeling.

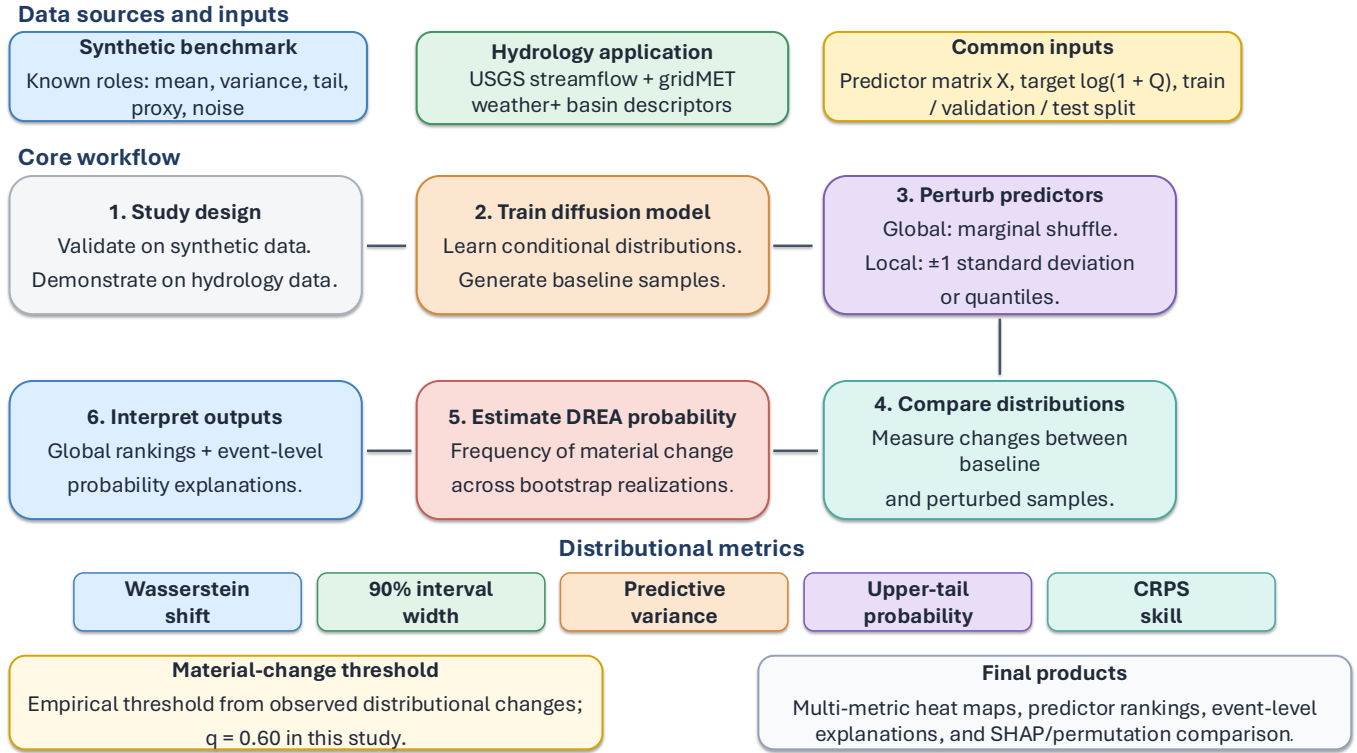


**Figure 1.** Overview of the Distributional Reliability Explanation Attribution (DREA) framework.

## 2.1 Study design and datasets

This study evaluates DREA in two settings. First, a controlled synthetic benchmark is used because the true role of each predictor is known. Second, a hydrological forecasting application is used to test whether the same framework can be applied to real diffusion-generated streamflow distributions. The synthetic benchmark is the main validation experiment; the hydrology case study serves as a real-world demonstration.

### 2.1.1 Synthetic benchmark

Let the predictor vector be

$$X = \{x_{\text{mean}}, x_{\text{var}}, x_{\text{tail}}, x_{\text{proxy}}, x_{\text{noise}}\}$$

The predictors were generated as independent standard normal variables, except for the proxy variable:

$$x_{\text{proxy}} = x_{\text{mean}} + \eta \quad 2$$

where  $\eta \sim \mathcal{N}(0, 0.35^2)$ .

Thus,  $x_{\text{proxy}}$  is correlated with the true mean driver but is not itself the primary distributional driver.

The conditional mean was defined as

$$\mu(X) = 1.6x_{\text{mean}} \quad 3$$

The predictive spread was controlled by  $x_{\text{var}}$ :

$$\sigma(X) = 0.15 + 1.45 \cdot \text{sigmoid}(2.8x_{\text{var}}) \quad 4$$

where

$$\text{sigmoid}(z) = \frac{1}{1 + \exp(-z)} \quad 5$$

The base response was generated as

$$y_{\text{base}} = \mu(X) + \sigma(X)\epsilon \quad 6$$

Where  $\epsilon \sim \mathcal{N}(0,1)$ .

Upper-tail events were introduced through a Bernoulli process controlled by  $x_{\text{tail}}$ :

$$p_{\text{tail}} = 0.03 + 0.38 \cdot \text{sigmoid}(3.2x_{\text{tail}}) \quad 7$$

$z_{\text{tail}} \sim \text{Bernoulli}(p_{\text{tail}})$ .

The final response was

$$y = y_{\text{base}} + z_{\text{tail}}(2.7 + 0.8|\xi|) + \delta \quad 8$$

Where  $\xi \sim \mathcal{N}(0,1), \delta \sim \mathcal{N}(0, 0.04^2)$ .

The variable  $\epsilon$  controls the main heteroscedastic spread of the predictive distribution,  $\xi$  controls the stochastic magnitude of upper-tail events, and  $\delta$  represents a small independent residual noise term added to avoid unrealistically deterministic structure in the synthetic benchmark.

This design creates known roles such that:  $x_{\text{mean}}$  controls the mean,  $x_{\text{var}}$  controls uncertainty,  $x_{\text{tail}}$  controls upper-tail behavior,  $x_{\text{proxy}}$  is a correlated proxy, and  $x_{\text{noise}}$  is irrelevant. The variable  $x_{\text{noise}}$  was intentionally excluded from the data-generating process and therefore represents a purely irrelevant predictor with no direct influence on the predictive distribution. All synthetic variables were generated using a fixed random seed to ensure full reproducibility.

While this synthetic design prioritizes interpretable ground-truth roles over full replication of hydrological complexity, including strong temporal autocorrelation, multimodality, and spatial dependence, it provides a controlled environment for evaluating whether DREA recovers distinct distributional mechanisms.

### 2.1.2 Hydrological data

The real-data application used daily streamflow observations from United States Geological Survey stations in Maryland [21] with at least 95% temporal completeness over 2010–2024 (Figure S1). Daily discharge  $Q_t$  was transformed as

$$y_t = \log(1 + Q_t) \tag{10}$$

to reduce positive skewness and stabilize the variance of daily streamflow values prior to diffusion-model training.

Meteorological predictors were obtained from gridMET [22] and included daily precipitation, minimum temperature, and maximum temperature. To represent antecedent hydrological memory, precipitation was aggregated over rolling windows:

$$P_t^{(k)} = \sum_{r=0}^{k-1} P_{t-r} \tag{11}$$

where  $k \in \{3, 7, 14, 30\}$ . Temperature memory was represented using rolling means:

$$T_t^{(k)} = \frac{1}{k} \sum_{r=0}^{k-1} T_{t-r} \tag{12}$$

Static basin descriptors were also included:  $X_{\text{static}} = \{\text{basin area, basin perimeter}\}$ . No lagged discharge, month, season, or day-of-year variables were used, so the explanations focus on meteorological memory and basin descriptors rather than autoregressive discharge information.

## 2.2 Model Training and Prediction

### 2.2.1 Conditional diffusion forecasting model

For each sample  $i$ , the model receives a predictor vector  $X_i$  and generates a distribution of possible target values  $Y_i$ . In the hydrological application,  $X_i$  contains meteorological memory variables and basin descriptors for a given station-date, and  $Y_i$  is the transformed daily discharge,  $y_i = \log(1 + Q_i)$ . The conditional predictive distribution is written as  $P_\theta(Y_i | X_i)$ , where  $\theta$  denotes the trained diffusion-model parameters. In practice, this distribution is represented using  $S$  generated samples:

$$\hat{Y}_i = \{\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \dots, \hat{y}_i^{(S)}\} \quad 13$$

In the streamflow example, these generated samples represent multiple plausible daily discharge outcomes for the same station-date, conditional on the observed precipitation, temperature, and basin attributes. The diffusion model was trained by gradually adding Gaussian noise to the observed transformed discharge and learning to remove that noise conditional on  $X_i$ . For diffusion timestep  $t$ , the noisy target is

$$y_t = \sqrt{\bar{\alpha}_t} y_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \quad 14$$

where  $y_0$  is the observed transformed discharge,  $\epsilon \sim \mathcal{N}(0,1)$ , and  $\bar{\alpha}_t$  is the cumulative noise schedule. The denoising network learns to estimate the injected noise:  $\epsilon_\theta(y_t, t, X_i) \approx \epsilon$ .

The training objective is

$$\mathcal{L}(\theta) = \mathbb{E}[\|\epsilon - \epsilon_\theta(y_t, t, X_i)\|^2] \quad 15$$

After training, prediction proceeds by starting from random noise and repeatedly applying the learned denoising model while conditioning on  $X_i$ . Repeating this sampling process produces an ensemble of plausible discharge values for each station-date. The predictive median summarizes the central forecast, while the empirical 5th and 95th percentiles summarize the 90% predictive interval. These generated samples form the predictive distribution that DREA later perturbs and compares across distributional metrics.

### 2.2.2 Baseline comparison with SHAP and permutation importance

SHAP was used as the primary point-attribution benchmark because it is one of the most widely adopted feature-attribution methods in applied machine learning and environmental modeling. SHAP is based on Shapley values from cooperative game theory and assigns each predictor a contribution to a model output [7]. This makes it a strong representative baseline for conventional feature-importance explanation. However, SHAP is usually applied to point

predictions or expected model outputs, whereas DREA is designed to explain full predictive distributions generated by a diffusion model.

To provide a standard and computationally stable SHAP baseline, we trained Random Forest regressors on the same predictor-target data and computed TreeSHAP values [24]. Permutation importance [8] was also computed as a supplementary comparator. These baselines were not intended to explain the internal mechanics of the diffusion model or to compare Random Forests with diffusion models as forecasting architectures. Instead, they were used to compare two explanation paradigms under the same synthetic ground-truth setting: conventional point-attribution of model output versus DREA-based distributional reliability attribution. This comparison is defensible because the synthetic benchmark explicitly defines which predictors control mean behavior, variance, upper-tail behavior, proxy correlation, and irrelevant noise. Therefore, the relevant question is whether each explanation approach recovers these known roles, not which predictive model has superior accuracy.

As a same-architecture point-output robustness check, we also applied model-agnostic SHAP to the predictive mean of a conditional diffusion model trained with the same synthetic data-generating process and architecture used for DREA. The diffusion sampler was converted to a point-output function by averaging generated samples. This comparison does not replace DREA because SHAP explains the expected prediction, whereas DREA explains changes in the full generated predictive distribution.

### 2.3 Predictor perturbation

For each predictor  $x_j$ , DREA evaluates its influence by comparing the predictive distribution generated from the original input  $X$  with the distribution generated after perturbing only  $x_j$ . For global DREA, the primary perturbation strategy was marginal shuffle. Under this strategy, the observed values of  $x_j$  are replaced by values randomly drawn from the empirical marginal distribution of  $x_j$  in the training data:

$$X^{(j\text{-shuf})} = (x_1, \dots, \tilde{x}_j, \dots, x_p) \tag{16}$$

where  $\tilde{x}_j \sim \hat{P}_{\text{train}}(x_j)$ .

This perturbation removes the observed alignment between  $x_j$  and the target-generating structure while preserving the empirical scale and marginal distribution of the predictor. The diffusion model is then sampled using  $X^{(j\text{-shuf})}$ , producing the perturbed predictive samples

$$\hat{\mathcal{Y}}^{(j\text{-shuf})} = \{\hat{y}_{j,\text{shuf}}^{(1)}, \dots, \hat{y}_{j,\text{shuf}}^{(S)}\} \tag{17}$$

Thus, global DREA compares the original generated distribution:  $\hat{\mathcal{Y}} = \{\hat{y}^{(1)}, \dots, \hat{y}^{(S)}\}$  with the marginal-shuffle perturbed distribution:  $\hat{\mathcal{Y}}^{(j\text{-shuf})}$ . Additional perturbations were used for local and event-level directional analyses. For a specific event  $i$ , positive and negative one-standard-deviation perturbations were defined as

$$\begin{aligned} X_i^{(j,+)} &= (x_{i1}, \dots, x_{ij} + \sigma_j, \dots, x_{ip}), \\ X_i^{(j,-)} &= (x_{i1}, \dots, x_{ij} - \sigma_j, \dots, x_{ip}), \end{aligned}$$

where  $\sigma_j$  is the empirical standard deviation of predictor  $x_j$ . These directional perturbations answer local questions such as whether increasing or decreasing a predictor raises the median prediction, widens uncertainty, or increases upper-tail risk for a specific event.

Quantile-replacement perturbations were also evaluated as sensitivity checks, where  $x_j$  was replaced with empirical lower- or upper-quantile values from the training distribution.

## 2.4 Distributional comparison

After the baseline and perturbed predictive samples have been generated, DREA compares them using distributional metrics that capture different aspects of predictive behavior. This step is necessary because a predictor may influence the generated distribution in different ways: it may shift the center of the distribution, widen or narrow uncertainty intervals, change predictive variance, alter upper-tail risk, or affect probabilistic forecast skill. Therefore, rather than relying on a single importance score, DREA evaluates multiple complementary distributional changes. These changes are then converted into reliability probabilities that indicate how consistently each predictor alters each aspect of the predictive distribution.

## 2.5 Distributional Reliability Explanation Attribution

For predictor  $x_j$ , distributional metric  $k$ , and bootstrap world  $m$ , DREA computes the change between the original predictive samples and the perturbed predictive samples:

$$\Delta_{j,k}^{(m)} = d_k(\hat{\mathcal{Y}}^{(m)}, \hat{\mathcal{Y}}^{(j,m)}) \tag{18}$$

where  $d_k(\cdot, \cdot)$  is a distributional comparison metric. In this study,  $d_k$  includes Wasserstein distance, interval-width change, variance change, upper-tail probability change, and Continuous Ranked Probability Score (CRPS) change.

The DREA probability is defined as

$$\text{DREA}_{j,k} = \frac{1}{M} \sum_{m=1}^M \mathbb{I}(\Delta_{j,k}^{(m)} > \varepsilon_k) \quad 19$$

where  $M$  is the number of bootstrap realizations,  $\varepsilon_k$  is a material-change threshold for metric  $k$ , and  $\mathbb{I}(\cdot)$  is the indicator function.

For each distributional metric  $k$ , the material-change threshold  $\varepsilon_k$  was computed from the pooled empirical distribution of  $\Delta_{j,k}^{(m)}$  values across predictors and bootstrap realizations within the same experiment.

$$\varepsilon_k = Q_q(\{\Delta_{j,k}^{(m)}\}) \quad 20$$

Thus, all predictors were compared against a common metric-specific threshold. The primary analysis used  $q = 0.60$ , corresponding to a moderate empirical material-change threshold. This avoids selecting an arbitrary absolute threshold while preserving comparability across predictors within each DREA component. Sensitivity to the threshold choice was evaluated in the synthetic benchmark using  $q \in \{0.50, 0.60, 0.70, 0.80\}$  and is reported in the Supplementary Material.

In simple terms,  $\text{DREA}_{j,k}$  is the empirical probability that perturbing predictor  $x_j$  produces a meaningful change in distributional property  $k$ . Values close to 1 indicate that the predictor reliably controls that distributional property; values close to 0 indicate weak or unreliable distributional influence. DREA is evaluated using five complementary distributional metrics that quantify different aspects of predictive distribution perturbation: (1) global distributional shift, (2) uncertainty interval width, (3) distributional variance, (4) tail-risk behavior, and (5) probabilistic forecast skill.

### 2.5.1 Wasserstein-DREA (Global Distributional Shift)

Overall distributional shift was measured using the Wasserstein distance [23]:

$$W(P, Q) = \inf_{\gamma \in \Gamma(P, Q)} \int |u - v| d\gamma(u, v) \quad 21$$

where  $P$  and  $Q$  denote the original and perturbed predictive distributions,  $u$  and  $v$  are possible generated target values drawn from  $P$  and  $Q$ , respectively, and  $\Gamma(P, Q)$  is the set of all ways to match values from the two distributions. Intuitively, Wasserstein distance measures the minimum “transport cost” required to move the original predictive distribution into the perturbed predictive distribution. In DREA, a larger Wasserstein distance means that perturbing a predictor causes a stronger overall shift in the generated distribution.

### 2.5.2 Width-DREA (Uncertainty Interval Width)

Predictive interval width was computed as

$$\text{Width}_{90} = Q_{0.95} - Q_{0.05} \quad 22$$

The width change was

$$\Delta_{\text{width}} = | \text{Width}_{90}^{(j)} - \text{Width}_{90} | \quad 23$$

This asks whether a predictor widens or narrows uncertainty intervals.

### 2.5.3 Variance-DREA (Distributional Variance)

Predictive variance was

$$\text{Var}(Y) = \frac{1}{s-1} \sum_{s=1}^s (\hat{y}^{(s)} - \bar{y})^2 \quad 24$$

The variance change was

$$\Delta_{\text{var}} = | \text{Var}^{(j)}(Y) - \text{Var}(Y) | \quad 25$$

This evaluates whether a predictor changes the spread of possible generated outcomes.

### 2.5.4 Tail-DREA (Tail-risk Behavior)

Upper-tail probability was defined as

$$p_{\text{tail}} = P(Y > q_{0.95}) \quad 26$$

where  $q_{0.95}$  is a fixed upper-tail threshold defined from the observed target distribution used for evaluation. The threshold was held fixed when comparing baseline and perturbed generated distributions and was not recomputed separately for each perturbed predictive distribution. Because the threshold is fixed, Tail-DREA should be interpreted as attribution of fixed-threshold exceedance probability rather than as a purely tail-shape-only metric; predictors that strongly shift the distributional location can also change exceedance probability by moving generated samples above the fixed upper-tail threshold.

The tail change was

$$\Delta_{\text{tail}} = | p_{\text{tail}}^{(j)} - p_{\text{tail}} | \quad 27$$

This metric asks whether a predictor changes extreme-event probability.

Similarly lower-tail DREA can be formulated using  $P(Y < Q_{0.05})$ . In the present study, Tail-DREA was formulated using upper-tail exceedance probabilities because the hydrological application focused on high-flow and flood-risk behavior. However, the framework is directly extensible to lower-tail or two-sided extreme-event attribution.

### 2.5.5 CRPS-DREA (Probabilistic Forecast Skill)

Probabilistic forecast skill was measured using the CRPS [11]:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{I}(z \geq y))^2 dz \quad 28$$

where  $F(z)$  is the cumulative distribution function of the generated predictive distribution evaluated at value  $z$ ,  $y$  is the observed target value, and  $\mathbb{I}(z \geq y)$  is an indicator function equal to 1 when  $z$  exceeds the observed value and 0 otherwise. CRPS measures how closely the full predictive distribution matches the observed outcome, with smaller values indicating better probabilistic forecast skill.

The CRPS change was

$$\Delta_{\text{CRPS}} = \text{CRPS}^{(j)} - \text{CRPS} \quad 29$$

Because smaller CRPS indicates better probabilistic forecast quality, positive values of  $\Delta_{\text{CRPS}}$  indicate that perturbing a predictor degraded probabilistic skill. Thus, CRPS-DREA measures how reliably perturbing a predictor produces a material deterioration in the quality of the generated predictive distribution.

## 2.6 Global and event-level DREA

The preceding sections define DREA probabilities for different distributional metrics. These attribution probabilities can be computed at different levels of analysis. Global DREA evaluates how a predictor influences predictive behavior across all evaluation samples. Event-level DREA evaluates how a predictor influences the predictive distribution for a specific observation. This distinction provides both system-level explanations and local event-level explanations. Global DREA uses all test samples:

$$\Delta_{j,k}^{(m)} = d_k \left( \{\hat{\mathcal{Y}}_i\}_{i=1}^n, \{\hat{\mathcal{Y}}_i^{(j)}\}_{i=1}^n \right) \quad 30$$

The focus is to evaluate how reliably each predictor shapes the predictive behavior of the full system.

Event-level DREA applies the same idea to one event  $i$ :

$$\Delta_{i,j,k}^{(m)} = d_k \left( \hat{\mathcal{Y}}_i, \hat{\mathcal{Y}}_i^{(j)} \right) \quad 31$$

For a specific station-date event, the question becomes: how does perturbing one predictor change the predictive distribution?

For example, the probability that a perturbation increases predicted discharge is

$$P_{\text{up}}(i, j) = \frac{1}{M} \sum_{m=1}^M \mathbb{I} \left( \tilde{y}_{i,j}^{(m)} > \tilde{y}_i^{(m)} \right) \quad 32$$

where  $\tilde{y}$  denotes the predictive median.

Similarly, the probability that upper-tail risk increases is

$$P_{\text{tail-up}}(i, j) = \frac{1}{M} \sum_{m=1}^M \mathbb{I} \left( p_{\text{tail},i,j}^{(m)} > p_{\text{tail},i}^{(m)} \right)$$

## 2.7 Outputs and interpretation

The final DREA outputs are designed to support both global model interpretation and event-level explanation. At the global level, DREA produces a multi-metric attribution matrix in which each row corresponds to a predictor and each column corresponds to a distributional property, such as distributional shift, uncertainty width, variance, tail-risk behavior, or probabilistic forecast skill. This matrix can be visualized as heat maps or ranked bar plots to identify predictors that reliably control different aspects of the generated predictive distribution.

For the synthetic benchmark, the global DREA matrix is used to evaluate whether the framework recovers the known roles of the mean, uncertainty, tail, proxy, and noise variables. For hydrological application, the same output is used to identify meteorological and basin predictors that influence different probabilistic properties of streamflow predictions.

At the event level, DREA produces local probability-based explanations for individual observations. These include the probability that directional perturbation increases the predictive median, widens the predictive interval, increases predictive variance, or raises upper-tail risk. These outputs provide local explanations that are naturally aligned with generative forecasting because they describe how perturbation change generated distributions rather than only point predictions.

Together, these outputs allow DREA to report predictor influence in three complementary forms: multi-metric global attribution, comparison with conventional point-attribution baselines, and event-level probabilistic explanations for individual forecast cases

## 3 Results

### 3.1 Synthetic benchmark validation

The synthetic benchmark was designed to evaluate whether DREA could recover known distributional roles under controlled conditions. Because the data-generating mechanisms were explicitly specified, the benchmark provides a direct test of whether a probabilistic explainability framework correctly distinguishes predictors controlling mean behavior, uncertainty, upper-tail extremes, correlated proxy information, and irrelevant noise.

Figure 2 summarizes the global multi-metric DREA probabilities for the synthetic benchmark. The results show clear separation of the known predictor roles across the different distributional components. The mean driver exhibited consistently high DREA

probabilities across Wasserstein-DREA and CRPS-DREA, indicating that perturbations to this predictor reliably shifted the predictive distribution and altered probabilistic forecast skill. In contrast, the uncertainty driver,  $x_{\text{var}}$ , defined in Equation 4, exhibited the strongest Width-DREA and Variance-DREA responses, indicating that DREA successfully identified predictors controlling predictive spread and uncertainty amplification rather than only shifts in predictive location.

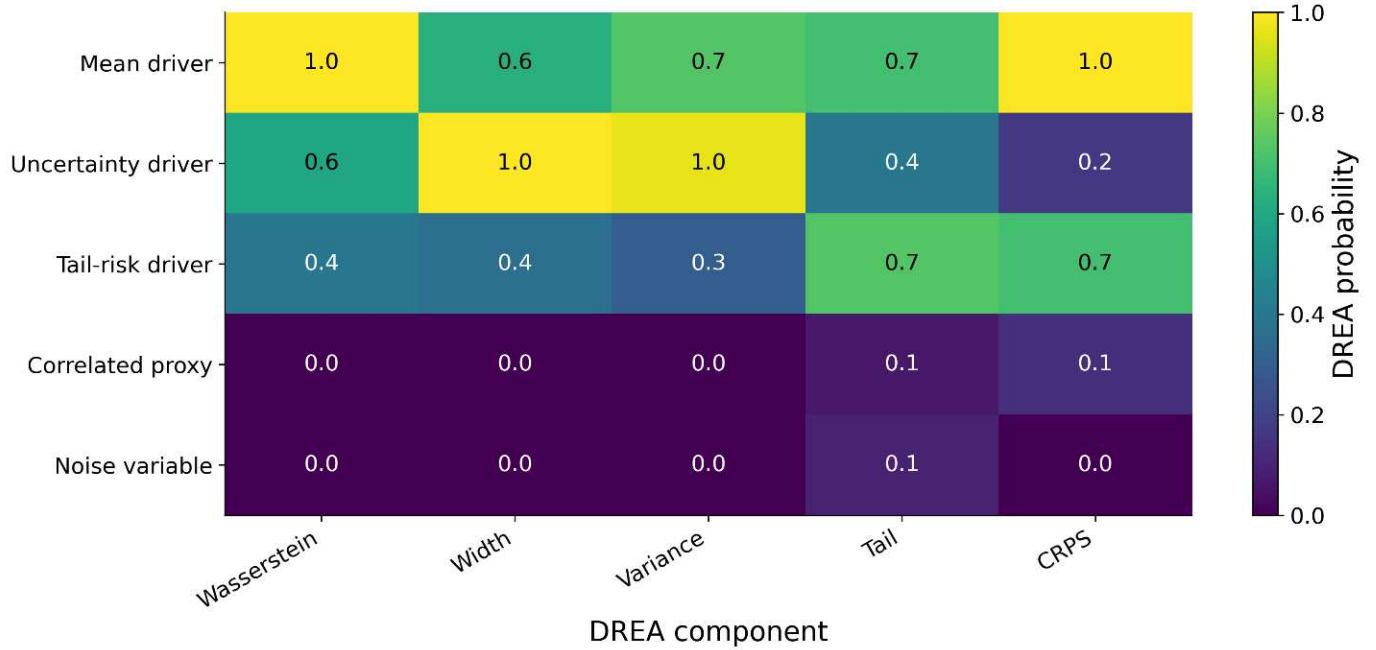


**Figure 2.** Synthetic benchmark heat map showing DREA recovery of known distributional roles

The tail driver produced elevated Tail-DREA and CRPS-DREA probabilities while showing weaker Wasserstein-, Width-, and Variance-DREA responses (Figure 2). This indicates that DREA captured its intended role as an upper-tail and probabilistic-skill driver rather than a primary location or uncertainty-width driver.

The correlated proxy experiment provides one of the clearest demonstrations of the framework’s value. Although the proxy variable was strongly correlated with the true mean driver, DREA assigned substantially lower distributional reliability probabilities to the proxy than to the known distributional driver. This indicates that DREA is not merely detecting predictive association; instead, it estimates whether perturbing a variable reliably alters the generated predictive distribution. Similarly, the irrelevant noise variable received near-zero DREA probabilities across nearly all metrics.

The comparison with SHAP highlights where point-attribution and distributional attribution agree and diverge (Figure 3). Both SHAP and DREA identified the mean driver as important, assigning it an importance value and reliability probability of 1.0, indicating agreement when

a predictor strongly controls prediction magnitude. However, SHAP assigned comparable importance to the correlated proxy and the tail-risk driver, whereas DREA assigned low reliability to the correlated proxy because perturbing it did not consistently alter the predictive distribution generated. SHAP also assigned low importance to the uncertainty driver,  $x_{\text{var}}$ , because this variable mainly controlled predictive spread rather than the average prediction. In contrast, DREA recovered this role through high Width-DREA and Variance-DREA probabilities. Permutation importance, reported in Table S2, showed even weaker sensitivity to non-mean distributional drivers, assigning near-zero importance to the uncertainty driver. These results indicate that conventional point-attribution baselines can recover dominant mean effects but may conflate correlated proxies and understate predictors that primarily control uncertainty or distributional behavior.

Supplementary robustness analyses further supported this interpretation. First, DREA probabilities were recomputed using material-change thresholds of  $q = 0.50, 0.60, 0.70$ , and  $0.80$  (Table S5). The expected driver structure was stable for the main distributional components: the Mean variable remained the top predictor for Wasserstein-DREA and CRPS-DREA across all thresholds, while the Uncertainty variable remained the top predictor for Width-DREA and Variance-DREA. Tail-DREA identified the Tail variable under moderate thresholds, while stricter thresholds increasingly emphasized the Mean variable, consistent with the fact that fixed-threshold exceedance probability can be affected by both tail-specific jumps and large location shifts. Second, SHAP applied directly to the diffusion predictive mean ranked the Mean variable highest and assigned low importance to the Uncertainty variable (Table S6). The Mean-proxy variable also received low diffusion-mean SHAP importance, indicating that proxy inflation was not present in this same-architecture point-output check. Therefore, the strongest evidence for DREA's added value is not universal superiority over SHAP on correlated proxies, but its ability to show what aspect of the predictive distribution each predictor drives. SHAP applied to the diffusion predictive mean provides a single point-output importance score, whereas DREA decomposes predictor influence into separate effects on distributional shift, uncertainty width, variance, tail-risk probability, and probabilistic skill.

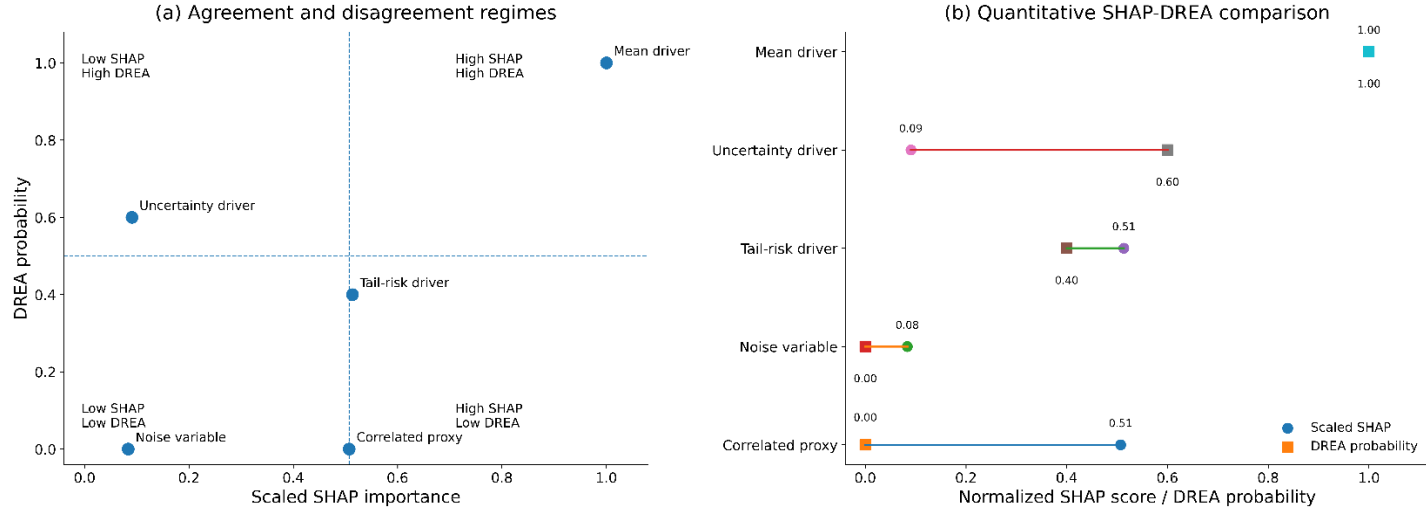


**Figure 3.** Comparison between Random-Forest SHAP importance and Wasserstein-DREA in the synthetic benchmark. (a) Quadrant interpretation showing agreement and disagreement regimes between association-based and distributional explanations. (b) Quantitative comparison of normalized SHAP scores and DREA probabilities across synthetic variables.

### 3.2 Real-world hydrological application

The generalized DREA framework was next applied to diffusion-generated streamflow predictions using daily meteorological and basin predictors from Maryland streamflow stations. Unlike the synthetic benchmark, the real-world experiment does not provide known ground-truth attribution structure. It therefore serves as an external application test of whether DREA yields interpretable distributional decompositions in a realistic forecasting setting.

Before applying DREA, we verified that the explained diffusion model achieved strong deterministic skill and calibrated probabilistic performance under station-blocked validation (Table S1). This ensured that DREA was applied to a functionally competent generative predictive model rather than to an inadequately trained forecasting system.

Figure 4 summarizes the hydrological multi-metric DREA probabilities. Basin perimeter showed a DREA probability of 1.0 across all five metrics, indicating that perturbing this variable produced material changes in every bootstrap replication under the selected thresholding scheme. This does not imply that basin perimeter is a perfect causal driver of streamflow; rather, it indicates that, within the trained diffusion model, basin perimeter acted as a highly stable structural predictor influencing distributional shift, uncertainty width, variance, upper-tail probability, and probabilistic skill. Basin area also showed strong DREA probabilities, especially for Wasserstein-DREA, Tail-DREA, and CRPS-DREA, suggesting that basin-scale descriptors carried broad distributional information in the hydrological application.

Meteorological memory variables displayed more differentiated attribution structures. Longer-memory variables, particularly 30-day maximum temperature, 30-day minimum temperature, and 30-day precipitation, showed stronger Tail-DREA and CRPS-DREA responses than most daily or 7-day predictors. In contrast, 3-day precipitation, 3-day maximum temperature, 14-day maximum temperature, and 14-day minimum temperature showed stronger Width-DREA or Variance-DREA responses while exhibiting weaker Wasserstein-DREA or Tail-DREA behavior. This suggests that some short- to intermediate-memory meteorological predictors primarily influenced predictive spread and uncertainty rather than the overall location of the discharge distribution.

Furthermore, the hydrological experiment demonstrates that DREA can separate multiple probabilistic mechanisms within a real diffusion-generated forecasting system. Rather than producing a single predictor ranking, the framework reveals how different variables influence distinct aspects of predictive behavior, including distributional shift, uncertainty amplification, and upper-tail discharge risk.

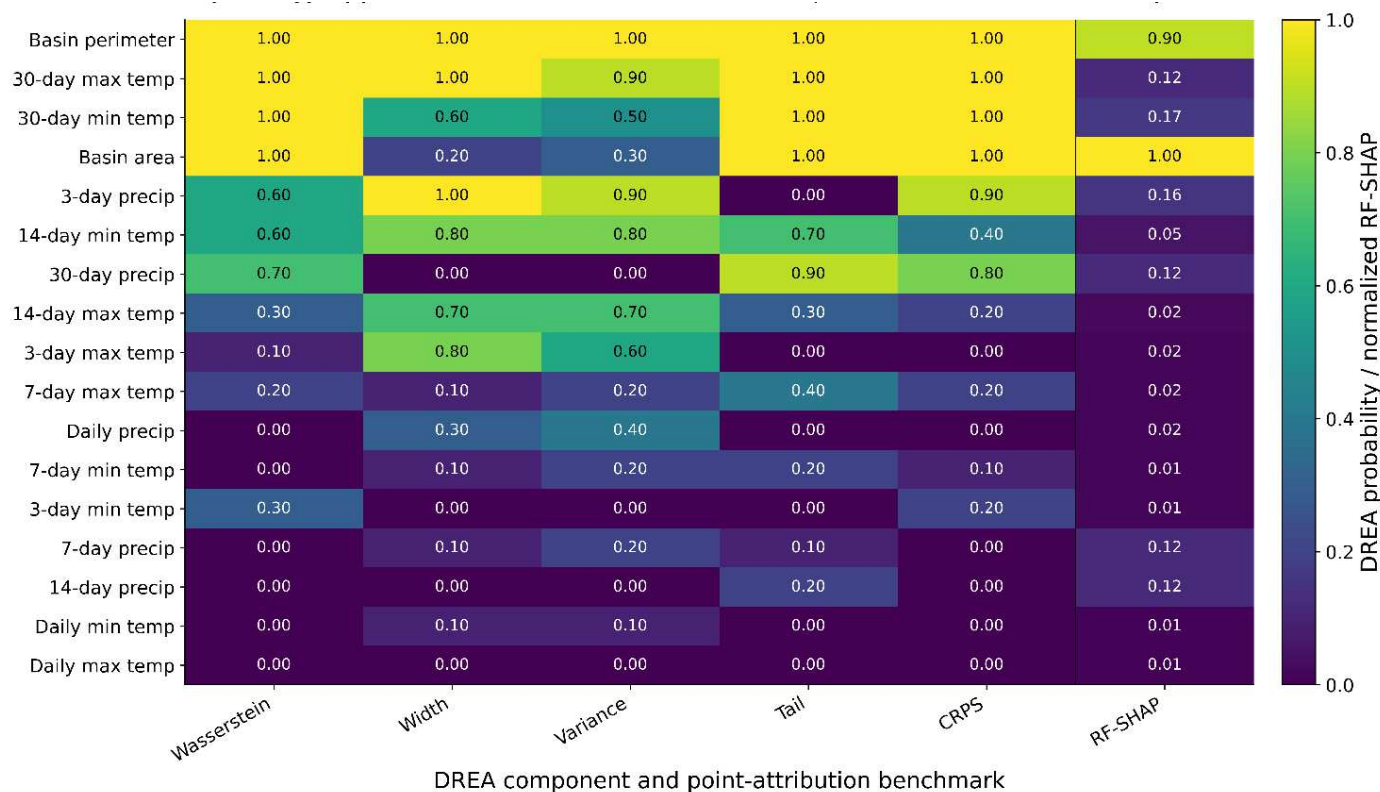


**Figure 4.** Hydrological DREA decomposition with RF-SHAP comparison. Rows show the top hydrological predictors ranked by mean DREA probability across distributional components. The first five columns show DREA probabilities. The final column shows Random Forest SHAP importance normalized to the range 0–1 and should therefore be interpreted as a scaled point-prediction importance score rather than a distributional reliability probability.

To provide a comparison with conventional point-attribution methods, SHAP values were also computed for the hydrological prediction model (Figure 4, Table S3). Both DREA and SHAP consistently identified basin-scale descriptors, particularly basin area and basin perimeter, as highly influential predictors. This agreement suggests that the strong basin-geometry signal is not solely an artifact of DREA thresholding; rather, basin area and basin perimeter encode stable structural information used by both the diffusion-based distributional attribution framework and the point-attribution baseline.

However, several meteorological memory variables exhibited substantially stronger DREA responses than SHAP importance scores. For example, 30-day precipitation, 30-day maximum temperature, and 30-day minimum temperature showed high Tail-DREA, Wasserstein-DREA, and CRPS-DREA probabilities despite low SHAP importance values. These results suggest that some predictors contribute primarily to predictive uncertainty, probabilistic skill, or extreme-event behavior rather than average prediction magnitude. Consequently, DREA provides complementary information to conventional point-attribution methods by identifying predictors that shape the predictive distribution even when their contribution to point predictions is relatively modest.

### **3.3 Event-level probabilistic explanations**

In addition to global attribution, DREA produced event-level probabilistic explanations for individual forecast events. For each selected station-date event, predictors were perturbed using directional interventions, and the diffusion model was repeatedly sampled to estimate how often the perturbation increased upper-tail discharge risk.

Figure 5 shows that event-level attribution can vary across forecast cases, even for predictors that are globally important. Basin perimeter had a DREA probability of 1.0 across all global metrics in Figure 4, and most selected events in Figure 5 also showed high basin-perimeter upper-tail response probabilities. However, one event showed a weak response, indicating that a globally reliable distributional predictor may not increase tail risk for every individual event under a +1 standard deviation perturbation. Similarly, basin area and 3-day precipitation showed strong responses for some events but weaker responses for others.

Notably, several meteorological memory variables that exhibited strong global Tail-DREA probabilities, including 30-day precipitation and 30-day temperature metrics, showed more heterogeneous event-level responses. This suggests that these predictors act as conditional tail-risk drivers whose influence depends on the local hydrometeorological state of individual forecast events, whereas basin-scale descriptors exhibited more stable event-level behavior.

These event-level results should be interpreted as model-based sensitivity patterns rather than causal hydrological mechanisms or site-specific operational predictions. The main contribution of the event-level analysis is to show that DREA can complement global attribution by providing local probabilistic explanations of how predictor perturbations alter generated tail-risk distributions. We did not compare event-level DREA directly with SHAP

because SHAP provides local contributions to point predictions, whereas event-level DREA estimates probabilities that perturbations change distributional properties such as upper-tail risk.

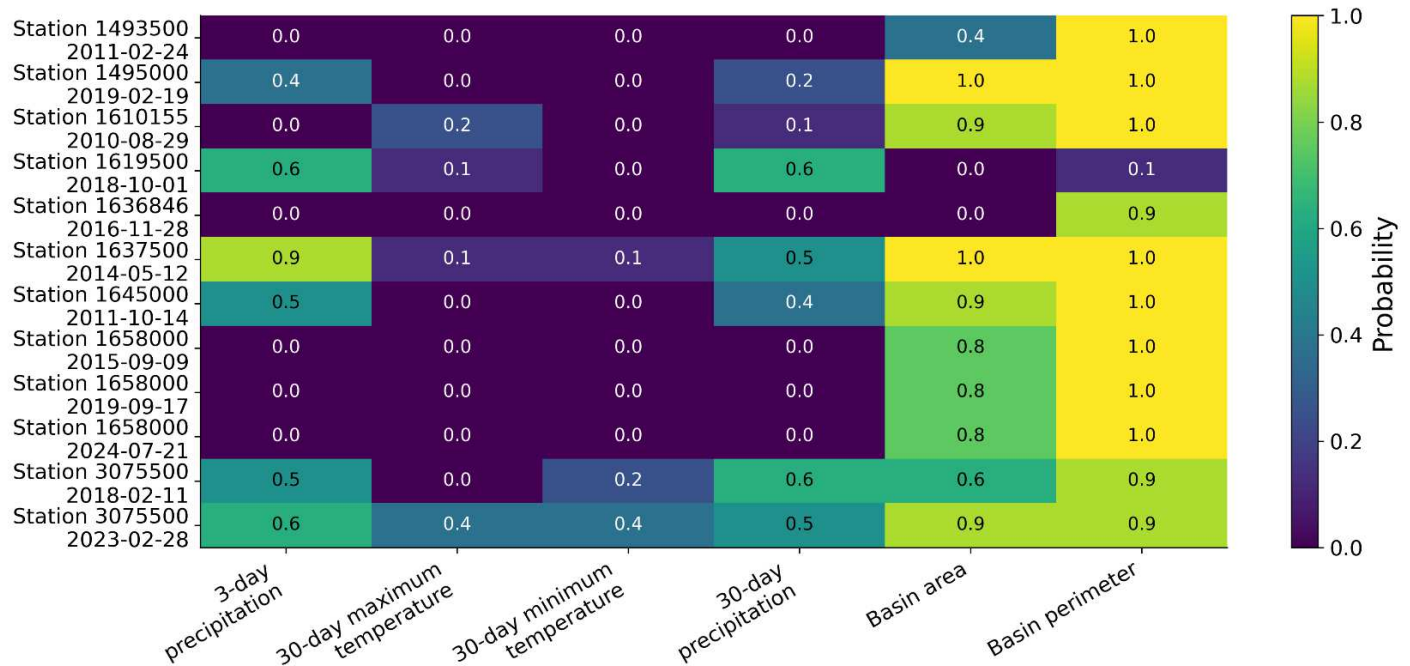


**Figure 5.** Event-level DREA analysis for upper-tail flow risk. Values represent the probability that a +1 standard deviation perturbation of a predictor increases upper-tail discharge risk for specific station-event combinations.

## 4. Discussion

### 4.1 Probabilistic distributional attribution for generative forecasting

Several previous studies have explored explainability in environmental and probabilistic forecasting using SHAP, integrated gradients, attention visualization, or permutation-based importance methods [25-28]. However, these approaches typically explain prediction magnitude or local prediction contribution rather than explicitly quantifying probabilistic distributional control [7-11]. DREA differs conceptually because it frames explanation as the probability that perturbing a predictor materially changes a generated predictive distribution. In this sense, the framework shifts explainability from deterministic feature contribution toward empirical distributional reliability.

The results show why this shift is useful for generative forecasting systems. In the synthetic benchmark, different predictors controlled different aspects of the predictive distribution. The mean driver mainly affected distributional location and probabilistic skill, the variance

driver mainly affected predictive width and variance, and the tail driver mainly affected upper-tail behavior. This confirms that predictor influence in probabilistic forecasting is not one-dimensional. A single importance score cannot clearly distinguish whether a predictor shifts the center of the distribution, widens uncertainty intervals, changes variance, alters probabilistic skill, or controls extreme-event probability [4, 14, 29].

The comparison with SHAP and permutation importance further supports this point. SHAP and permutation importance were informative for predictors associated with average prediction magnitude, but they are not designed to decompose predictor effects into separate distributional components such as uncertainty width, variance, and tail-risk behavior [15,30].

The correlated proxy result requires nuanced interpretation. In the Random-Forest SHAP baseline, the Mean-proxy variable received elevated importance because it was strongly correlated with the true Mean variable. In contrast, DREA assigned low distributional reliability to the Mean-proxy variable because perturbing it did not consistently alter the generated predictive distribution. However, same-architecture SHAP applied to the diffusion predictive mean also assigned low importance to the Mean-proxy variable. This indicates that proxy inflation is not inevitable for every SHAP implementation. Therefore, the main advantage of DREA is not simply that it suppresses correlated variables better than SHAP, but that it provides a multi-component distributional explanation that separates mean/location effects, uncertainty width, variance, tail-risk behavior, and probabilistic skill.

The same-architecture diffusion-mean SHAP analysis strengthens this distinction. When SHAP was applied directly to the predictive mean of the conditional diffusion model, the mean driver dominated the attribution ranking, whereas the uncertainty driver received low point-output importance. In contrast, DREA assigned high reliability to the uncertainty driver under Width-DREA and Variance-DREA. This shows that the contribution of DREA is not simply the use of a diffusion model, but the explicit attribution of changes in the full generated predictive distribution. In practical terms, SHAP can identify that a predictor is important for the diffusion predictive mean, but DREA indicates how it is important: whether it shifts the distribution, widens uncertainty, increases variance, changes upper-tail probability, or degrades probabilistic skill.

## **4.2 Real-world Applications: Beyond Hydrological Forecasting**

The hydrological application showed that DREA can be applied beyond controlled synthetic data to real diffusion-generated forecasts. In the streamflow case, DREA separated predictors associated with distributional shift, uncertainty amplification, variance control, probabilistic skill, and upper-tail discharge risk. This is useful because hydrological predictors do not necessarily influence forecasts in the same way. Basin descriptors may shape broad discharge distributions [31], antecedent precipitation and wetness conditions may influence

flood-response and upper-tail behavior [32], and temperature-memory variables may affect distributional state through snowmelt, rain-on-snow processes and evapotranspiration [33,34]. A conventional point-attribution ranking would make these mechanisms harder to separate.

This distributional decomposition has direct relevance for downstream decision support. In flood forecasting, for example, a predictor that mainly increases upper-tail probability may be more relevant for early warning than a predictor that only shifts the median forecast. In reservoir operation or water-resource planning, predictors associated with wider uncertainty intervals may indicate cases where conservative decisions or additional monitoring are needed. In energy forecasting, similar DREA components could help distinguish predictors that affect expected demand from those that amplify price or load uncertainty. In biomedical risk prediction, the framework could separate variables that increase average predicted risk from variables that widen prognostic uncertainty or increase extreme-risk probability. Thus, DREA can support more informed decisions by identifying not only which predictors matter, but how they matter for uncertainty-sensitive outcomes.

The event-level results further extend the framework from global interpretation to local probabilistic explanation. Instead of asking only which predictors are generally important, event-level DREA asks how a perturbation changes the predictive distribution for a specific station-date event. This allows statements such as whether increasing antecedent precipitation makes high-flow outcomes more likely, whether uncertainty widens, or whether the predictive median increases across diffusion-generated realizations. Such explanations are naturally aligned with generative models because they interpret changes in distributions rather than only changes in point predictions.

Overall, the results suggest that probabilistic distributional attribution may provide a more appropriate explanation framework for diffusion and generative forecasting systems than conventional point-attribution methods alone. As generative models become increasingly common in environmental forecasting [12,35], infrastructure planning [36], energy systems [37,38], biomedical prediction [39,40], and other uncertainty-sensitive applications [41,42], methods capable of explaining predictive distributions rather than only point predictions will become increasingly important. As a modular component of expert and intelligent decision-support systems, DREA can enhance trustworthiness by providing probabilistic explanations that are aligned with the generative nature of diffusion-based forecasts.

### **4.3 Limitations**

The main practical limitation of DREA is computational cost. The method requires repeated perturbations, bootstrap realizations, and diffusion sampling, making it more expensive than conventional point-attribution methods. In addition, some implementation choices, including the material-change threshold, number of perturbation realizations, number of

generated samples, and perturbation strategy, are user-defined and may influence attribution probabilities. In this study, these choices were made transparently and consistently across experiments (Table S4), with the primary material-change threshold defined empirically using the 60th percentile of observed distributional shifts. Synthetic threshold-sensitivity analysis using  $q=0.50, 0.60, 0.70$ , and  $0.80$  is reported in Table S5. Future work should extend similar threshold-sensitivity analyses to larger real-world applications as computational resources permit.

## 5. Conclusion and Future Work

This study introduced Distributional Reliability Explanation Attribution (DREA), a probabilistic explainability framework for diffusion-based forecasting systems. Unlike conventional point-attribution methods, DREA estimates how reliably predictor perturbations alter different properties of generated predictive distributions, including distributional shift, uncertainty width, predictive variance, upper-tail behavior, and probabilistic forecast skill.

A controlled synthetic benchmark with known probabilistic generating mechanisms demonstrated that DREA successfully recovers distinct distributional roles associated with mean behavior, uncertainty amplification, and extreme-event control while assigning lower reliability to correlated proxy variables and near-zero reliability to irrelevant noise. Comparisons with Random-Forest SHAP, diffusion-mean SHAP, and permutation importance showed that point-output attribution methods can recover dominant mean effects, but they provide limited separation of predictors that primarily influence predictive uncertainty, variance, or tail-specific behavior. The correlated-proxy result was method-dependent: Random-Forest SHAP assigned elevated importance to the Mean-proxy variable, whereas diffusion-mean SHAP and DREA both assigned it low importance or reliability.

Supplementary same-architecture analysis using SHAP on the diffusion predictive mean confirmed that point-output attribution primarily emphasized predictors controlling expected prediction magnitude, while DREA recovered predictors controlling uncertainty width and variance through distributional components.

The framework was additionally demonstrated on diffusion-generated hydrological forecasts using Maryland streamflow, meteorological, and basin predictors. The real-world application showed that DREA can decompose predictor influence into interpretable probabilistic components and can provide both global and event-level explanations for generated predictive distributions.

Overall, the results suggest that explainability in generative forecasting systems should extend beyond deterministic feature importance toward probabilistic distributional

attribution. DREA provides one possible framework for this transition by aligning explainability with the probabilistic nature of diffusion-generated forecasts.

Future work will extend DREA in three directions. First, additional evaluation metrics can be incorporated to capture lower-tail risk, two-sided extremes, calibration error, interval score, distributional sharpness, and decision-weighted losses. This would allow DREA to be adapted to applications where different parts of the predictive distribution matter, such as drought risk, flood risk, electricity price spikes, or clinical deterioration.

Second, future implementations should develop an automated DREA workflow that includes perturbation selection, sensitivity analysis, uncertainty calibration, metric selection, and standardized visualization. Such a workflow could produce multi-metric heat maps, global predictor rankings, event-level explanation panels, and comparison summaries with conventional point-attribution methods. This would make DREA easier to apply across domains and support more transparent interpretation of generative forecasting systems.

## **Declarations**

### **CRedit authorship contribution statement**

Chibuike Chiedozie Ibebuchi: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Validation ; Writing – original draft, Writing – review and editing.

Qunying Huang: Writing – review and editing; Validation

### **Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### **Data availability**

Synthetic benchmark code, DREA implementation, configuration files, and documentation are provided in the accompanying public software archive. USGS and gridMET inputs used for the hydrological application are publicly available from the cited sources.

### **Code availability**

The reproducibility code for DREA is provided as a public software archive: “*Ibebuchi CC (2026). Distributional Reliability Explanation Attribution for Diffusion Models. Zenodo. <https://doi.org/10.5281/zenodo.20579923>*”. The archive includes a generic implementation of the Distributional Reliability Explanation Attribution framework, the controlled synthetic

benchmark, and a command-line workflow for applying DREA to user-supplied CSV datasets, environment files, and documentation.

## **Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## **References**

- [1] De León Pérez, D., Salazar-Galán, S., & Francés, F. (2025). Beyond deterministic forecasts: A scoping review of probabilistic uncertainty quantification in short-to-seasonal hydrological prediction. *Water*, 17(20), 2932.
- [2] Ibebuchi, C. C., & Abu, I. O. (2025). Probabilistic flood susceptibility mapping using explainable AI for the Western United States. *Environmental Research Communications*, 7(10), 105008.
- [3] Newlands, N. K., Zamar, D. S., Kouadio, L. A., Zhang, Y., Chipanshi, A., Potgieter, A., ... & Hill, H. S. (2014). An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. *Frontiers in Environmental Science*, 2, 17.
- [4] Ou, Z., Nai, C., Pan, B., Zheng, Y., Shen, C., Jiang, P., ... & Pan, M. (2025). Probabilistic diffusion models advance extreme flood forecasting. *Geophysical Research Letters*, 52(15), e2025GL115705.
- [5] Wang, C., Guo, L., Fu, Z., Yang, S., Cheng, H., Kot, A. C., & Wen, B. (2025). Reconciling stochastic and deterministic strategies for zero-shot image restoration using diffusion model in dual. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 23207-23216).
- [6] Kumarage, P. M., & Saarela, M. (2026). Explainable Generative AI: A Two-Stage Review of Existing Techniques and Future Research Directions. *AI*, 7(1), 31.
- [7] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- [8] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- [9] Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR.
- [10] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- [11] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477), 359-378.

- [12] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840-6851.
- [13] Rasul, K., Seward, C., Schuster, I., & Vollgraf, R. (2021, July). Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International conference on machine learning* (pp. 8857-8868). PMLR.
- [14] Ibebuchi, C. C. (2025). Uncertainty in machine learning feature importance for climate science: a comparative analysis of SHAP, PDP, and gain-based methods. *Theoretical and Applied Climatology*, 156(9), 476.
- [15] Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298, 103502.
- [16] Janzing, D., Minorics, L., & Blöbaum, P. (2020, June). Feature relevance quantification in explainable AI: A causal problem. In *International Conference on artificial intelligence and statistics* (pp. 2907-2916). PMLR.
- [17] Hooker, S., Erhan, D., Kindermans, P. J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.
- [18] Marjanovic, S., Augenstein, I., & Lioma, C. (2024, August). Investigating the impact of model instability on explanations and uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11854-11879).
- [19] Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems*, 34, 9391-9404.
- [20] Löfström, H., Löfström, T., Johansson, U., & Sönströd, C. (2024). Calibrated explanations: With uncertainty information and counterfactuals. *Expert Systems with Applications*, 246, 123154.
- [21] U.S. Geological Survey. (2026). USGS Water Data for the Nation: U.S. Geological Survey National Water Information System database. Retrieved May 25, 2026, from <https://doi.org/10.5066/F7P55KJN>
- [22] Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *International journal of climatology*, 33(1), 121-131.
- [23] Villani, C. (2009). *Optimal Transport: Old and New*. Springer
- [24] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1), 56-67.
- [25] Başağaoğlu, H., Chakraborty, D., Lago, C. D., Gutierrez, L., Şahinli, M. A., Giacomoni, M., ... & Şengör, S. S. (2022). A review on interpretable and explainable artificial intelligence in hydroclimatic applications. *Water*, 14(8), 1230.

- [26] Liao, W., Fang, J., Ye, L., Bak-Jensen, B., Yang, Z., & Porte-Agel, F. (2024). Can we trust explainable artificial intelligence in wind power forecasting?. *Applied Energy*, 376, 124273.
- [27] Ozupek, E., Teke, A., Celik, N., & Kavzoglu, T. (2025). Explainable artificial intelligence to explore the intrinsic characteristics of climatic parameters governing meteorological drought forecasting: opening the black box. *Stochastic Environmental Research and Risk Assessment*, 39(8), 3201-3222.
- [28] Wegener, C., & Ibebuchi, C. C. (2025). Application of xgboost in disentangling the fingerprints of global warming and decadal climate modes on seasonal precipitation trends in ohio. *International Journal of Climatology*, 45(8), e8829.
- [29] Friederichs, P., & Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7), 579-594.
- [30] Verdinelli, I., & Wasserman, L. (2024). Feature importance: A closer look at shapley values and loco. *Statistical Science*, 39(4), 623-636.
- [31] Karalis, S., Karymbalis, E., Valkanou, K., Chalkias, C., Katsafados, P., Kalogeropoulos, K., ... & Bofilios, A. (2014). Assessment of the relationships among catchments' morphometric parameters and hydrologic indices. *International Journal of Geosciences*, 5(13), 1571-1583.
- [32] Massari, C., Pellet, V., Tramblay, Y., Crow, W. T., Gründemann, G. J., Hascoetf, T., ... & Marra, F. (2023). On the relation between antecedent basin conditions and runoff coefficient for European floods. *Journal of Hydrology*, 625, 130012.
- [33] Haleakala, K., Brandt, W. T., Hatchett, B. J., Li, D., Lettenmaier, D. P., & Gebremichael, M. (2023). Watershed memory amplified the Oroville rain-on-snow flood of February 2017. *PNAS nexus*, 2(1), pgac295.
- [34] Musselman, K. N., Clark, M. P., Liu, C., Ikeda, K., & Rasmussen, R. (2017). Slower snowmelt in a warmer world. *Nature Climate Change*, 7(3), 214-219.
- [35] Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., ... & Mohamed, S. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878), 672-677.
- [36] Wang, Q., Liang, Y., Zheng, Y., Xu, K., Zhao, J., & Wang, S. (2025). Generative AI for urban planning: Synthesizing satellite imagery via diffusion models. *Computers, Environment and Urban Systems*, 122, 102339.
- [37] Lu, X., Feng, Z., Sun, J., Lou, J., Wu, C., Bao, W., & Li, J. (2025, April). Generative diffusion model-based energy management in networked energy systems. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
- [38] Leibowicz, B. D., Krey, V., & Grubler, A. (2016). Representing spatial technology diffusion in an energy system optimization model. *Technological Forecasting and Social Change*, 103, 350-363.

- [39] Luo, J., Yang, L., Liu, Y., Hu, C., Wang, G., Yang, Y., ... & Zhou, X. (2025). Review of diffusion models and its applications in biomedical informatics. *BMC Medical Informatics and Decision Making*, 25(1), 390.
- [40] Guo, Z., Liu, J., Wang, Y., Chen, M., Wang, D., Xu, D., & Cheng, J. (2024). Diffusion models in bioinformatics and computational biology. *Nature reviews bioengineering*, 2(2), 136-154.
- [41] Bose, S., Dutta, A., Nag, S., Zhang, J., Li, J., Karydis, K., & Roy-Chowdhury, A. K. (2025). Uncertainty-Aware Diffusion-Guided Refinement of 3D Scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 28271-28281).
- [42] Alekya, G., Nagalakshmi, M., Verma, O. D., Ramkumar, A., & Kumar, P. (2025, November). Denoising Diffusion Probabilistic Models with Federated Learning for Privacy-Preserving Wind Power Forecasting. In *2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)* (pp. 1-7). IEEE.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryfiles.docx](#)