MANIFOLD CONSTRAINED CONFORMAL PREDICTION FOR SPATIAL DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a conformal prediction method for spatial distributions observed through finite samples of events, such as tropical cyclone genesis and earthquake locations. Each observation represents an underlying spatial distribution, and uncertainty concerns how this distribution varies across an observation period. Our approach scores empirical measures using spherical sliced Wasserstein distance, then constrains the resulting prediction set to lie near the training data manifold. We derive a finite sample coverage lower bound for future empirical measures using a data adaptive manifold tolerance. An allowance for sampling error extends this guarantee to the underlying spatial distribution. Because the resulting set is not analytically tractable, we introduce a modified flow based sampling procedure that represents candidate distributions by empirical density and produces an ensemble over spatial distributions. Numerical experiments assess empirical measure coverage on synthetic data and compare energy distance and manifold distance on synthetic processes, tropical cyclone genesis, and earthquake occurrences.

1 INTRODUCTION

Natural disasters are among the most consequential and difficult to predict phenomena in Geoscience. Tropical cyclones, earthquakes, and wildfires collectively account for trillions of dollars in economic losses and hundreds of thousands of deaths each year (Merz et al., 2020; Beven et al., 2018). These hazards manifest as discrete, spatially organized configurations of events rather than as continuous fields. Some examples are, specifically, the genesis locations of tropical cyclones in a given season, and epicenters of earthquake sequences. The generative density, positions, and spatial arrangement of these events are all meaningful, and the differences are particularly consequential for compound and extreme hazards, where dependence structure and spatial spread drive the largest risks (Davison & Huser, 2015; Zscheischler et al., 2018; Abbaszadeh et al., 2022). Uncertainty quantification for these phenomena cannot be reduced to scalar or low dimensional summaries.

Modern forecasting systems increasingly produce distributions over possible event configurations, either through probabilistic models or stochastic ensembles (Leutbecher & Palmer, 2008; Jordan et al., 2011; Zhou et al., 2022; Mukherjee et al., 2025). Figure 1 illustrates the spatial structure of the tropical cyclone genesis and earthquake data considered in this work. However, predictive distributions obtained from complex forecasting systems are not generally guaranteed to be calibrated. This motivates a post-hoc procedure that can be applied to arbitrary forecasting models while providing distribution-free coverage guarantees.

Conformal prediction (Vovk et al., 2005) provides exactly this guarantee under exchangeability (Shafer & Vovk, 2008). The prediction region is a sublevel set of a conformity score $S(y, \hat{y})$, and the coverage guarantee $\mathbb{P}(Y_{n+1} \in C_\alpha) \geq 1 - \alpha$ follows without distributional assumptions on the model or the data. The standard approach defines S as a norm of the residual $r = y - \hat{y}$, which is not well defined since subtraction has no natural meaning, and scalar summaries such as mean location or total count can be identical for densities with entirely different spatial organization.

We propose instead to carry out the comparison at the level of probability measures. Representing a spatial distribution y as its empirical measure μ_y maps the conformity scoring problem onto distributional comparison, for which optimal transport provides the appropriate geometry (Villani, 2009; Peyré & Cuturi, 2019). A conformity score based on W_2 produces prediction regions that are metric balls in $(P_2(\mathbb{S}^2), W_2)$, i.e. the set of all distributions within transport cost τ_α of the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

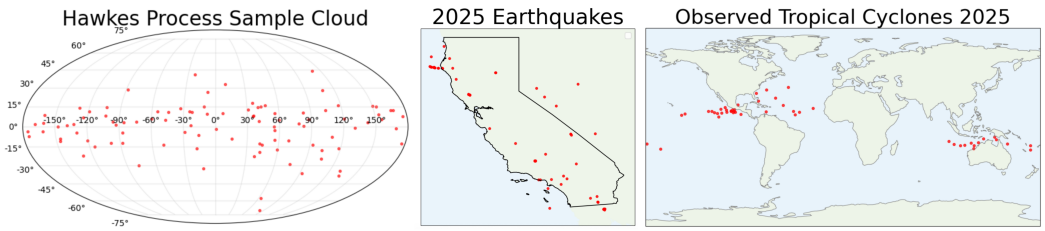

Figure 1: Examples of densities and point clouds on the sphere. *Right:* tropical cyclone genesis locations from one season (Gahtan et al., 2026), concentrated in tropical oceanic basins. *Center:* significant earthquake epicenters from the NGDC/WDS database (U.S. Geological Survey, 2026), tracing the global fault network. *Left:* a simulated calibration ensemble from a Hawkes process on the sphere (Section G). Each panel is a empirical density on $\mathbb{S}^2$.

predicted measure. The spherical sliced Wasserstein distance (Bonet et al., 2023) provides a tractable approximation at $O(n \log n)$ cost. The idea of using OT distances as conformity scores has been explored concurrently for multivariate Euclidean outputs (Thurin et al., 2025; Klein et al., 2025) and graph valued outputs (Melo et al., 2026); the present work addresses the space of probability measures on $\mathbb{S}^2$, where spherical density support requires different treatment.

Spatially adaptive conformal methods (Tibshirani et al., 2019; Barber et al., 2023) conformalize scalar residuals at individual locations and do not extend to the setting where the entire time interval point cloud configuration which is the prediction target. Bayesian spatial point process posteriors (Baddeley et al., 2015) provide an alternative source of uncertainty sets but require specifying a parametric generative model and do not offer post hoc coverage guarantees for an arbitrary forecasting system. Within the conformal literature, the concurrent OT based methods of (Thurin et al., 2025) and (Klein et al., 2025) operate on multivariate Euclidean outputs and do not accommodate the geodesic geometry of $\mathbb{S}^2$ densities; (Melo et al., 2026) use Gromov–Wasserstein distance for graphs, a structurally different object from empirical measures on the sphere. Osama et al. (Osama et al., 2019) construct spatially varying predictive intensity intervals for event counts after discretizing the spatial domain, whereas our goal is to construct regions for the full spatial probability distribution under a distributional transport score. Separately, Wasserstein regularized conformal methods have been developed to control coverage degradation under general train test distribution shifts (Xu et al., 2025). Such methods modify conformal calibration to accommodate shifts and are complementary to the present problem. Consequently, the natural comparison class for our method is HDR style region estimators applied to the forecast probability grid, which represent the current practical standard for spatial hazard uncertainty quantification.

Prediction sets defined as balls in $P_2(\mathbb{S}^2)$ are, however, far too vast to be useful (Villani, 2009). Without additional structure they contain physically implausible configurations, and unconstrained gradient flows in this space will generate off manifold samples even when the target distance is correct (Ambrosio et al., 2005). Building on the flow based conformal predictive distribution framework of (Harris, 2026), we introduce a constrained gradient flow that simultaneously targets the SSW prediction set boundary and the data manifold, producing an ensemble of boundary configurations that are both calibrated and geophysically coherent. Experiments show that the resulting regions achieve nominal coverage, score well under proper distributional scoring rules, and concentrate on physically plausible configurations in a way the unconstrained ambient ball does not.

2 Background

2.1 Conformal Prediction

Conformal prediction (Vovk et al., 2005) constructs prediction regions with finite sample coverage guarantees under exchangeability. Given calibration pairs $\{(X_i, Y_i)\}_{i=1}^n$ with responses in a space $\mathcal{Z}$ and a model f fitted on independent training data, a conformity score $S(Y, f(X))$ measures the discrepancy between the observed target and the model prediction. The calibration scores $S_i = S(Y_i, f(X_i))$ define the empirical $(1-\alpha)$-quantile $\tau_\alpha = S_{(\lceil (n+1)(1-\alpha) \rceil)}$, with $S_{(n+1)} = \infty$,

and the split conformal prediction region for a new test covariate X_{n+1} is

$$C_\alpha(X_{n+1}) = \{Y \in \mathcal{Z} : S(Y, f(X_{n+1})) \leq \tau_\alpha\}. \tag{1}$$

Under exchangeability, $\mathbb{P}(Y_{n+1} \in C_\alpha(X_{n+1})) \geq 1 - \alpha$. The framework extends to covariate shift (Tibshirani et al., 2019), distribution drift (Barber et al., 2023), functional data (Lei et al., 2015; Diquigiovanni et al., 2022), and flow based predictive distributions (Harris, 2026). The choice of conformity score determines the shape of the prediction region and how well it adapts to the problem geometry.

2.2 Wasserstein Distances on the Sphere

Let $P_2(\mathcal{Z})$ denote the space of Borel probability measures on a metric space $\mathcal{Z}$ with finite second moments. Optimal transport defines a geometry on $P_2(\mathcal{Z})$ by measuring the minimum cost of rearranging the mass of one measure to match another (Villani, 2009). For measures $\mu, \nu \in P_2(\mathcal{Z})$, the p-Wasserstein distance is

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{Z}\times\mathcal{Z}} d(x, y)^p \, d\pi(x, y) \right)^{1/p}, \tag{2}$$

where $\Pi(\mu, \nu)$ is the set of couplings with marginals μ and ν and d is the metric on $\mathcal{Z}$. For $\mathcal{Z} = \mathbb{S}^2$ the cost $d(x, y)$ is the geodesic (great circle) distance. For $p = 2$, the space $(P_2(\mathcal{Z}), W_2)$ is itself a metric space, and $W_2(\mu_n, \mu) \to 0$ if and only if $\mu_n \to \mu$ weakly and the second moments converge (Villani, 2009). A conformity score based on W_2 therefore defines prediction regions that are metric balls in $(P_2(\mathbb{S}^2), W_2)$: the set of all distributions within transport cost τ_α of the predicted measure.

Computing Wasserstein distance between discrete measures requires solving a transport problem. Spherical sliced Wasserstein distance (Bonet et al., 2023) adapts this construction by projecting onto great circles. For an orthonormal two-frame $U \in V_{3,2}$, the projection is $P^U(x) = U^\top x / \|U^\top x\|$, with a fixed value on $\mathbb{S}^1$ assigned when the denominator is zero. Throughout this paper we use L frames drawn uniformly and independently of the data, then held fixed across all evaluations:

$$\mathrm{SSW}_2^2(\mu, \nu) = \frac{1}{L} \sum_{\ell=1}^{L} W_2^2(P^{U_\ell}_{\#}\mu, P^{U_\ell}_{\#}\nu). \tag{3}$$

Here W_2 uses geodesic distance on the circle. This finite average approximates the integral over $V_{3,2}$ in the population definition. We retain the notation SSW_2 for its positive square root and use SSW_2^2 only for the squared form. The finite average is a pseudometric: it is symmetric and satisfies the triangle inequality, but distinct measures can have identical projections. The coverage arguments below require the triangle inequality, not separation of distinct measures.

3 Method

We now develop the conformal framework for spatial distributions observed through event samples. We first define the SSW score and the ambient region for empirical measures, then introduce an empirical surrogate for the data manifold. We give coverage bounds for both the empirical measure and the underlying distribution under the data manifold constraint.

3.1 Ambient Conformal Region

Let $P_i \in P_2(\mathbb{S}^2)$ be the underlying spatial probability measure for $i \in 1, ..., n$, with covariate $X_i \in \mathcal{X}$. We define $P_i(A) = \int_A p_i \, d\sigma$, for a geographic region $A \subseteq \mathbb{S}^2$, assuming P_i has a density p_i with respect to surface area. We observe a finite event sample $Y_i = \{Z_{ij}\}_{j=1}^{m_i}$ and represent it by

$$\widehat{P}_i = \mu_{Y_i} = \frac{1}{m_i} \sum_{j=1}^{m_i} \delta_{Z_{ij}}. \tag{4}$$

where each Z_{ij} is a specific hazard event location in the set Y_i. In general, we allow the sample size $m_i \geq 1$ to vary across observations $Y_1, Y_2, ..., Y_n$ since event counts enter only through the

empirical representation of the spatial probability measure. Let $\{(X_i^0, Y_i^0)\}_{i=1}^{n_0}$, $\{(X_i, Y_i)\}_{i=1}^{n}$, and (X_{n+1}, Y_{n+1}) be training, calibration, and test observations sampled i.i.d. from a common distribution. The prediction model $f : \mathcal{X} \to P_2(\mathbb{S}^2)$ is fit on the training pairs and then held fixed for calibration and testing. Given any new candidate measure μ, we define the conformal score as

$$S(\mu, f(X)) = \mathrm{SSW}_2(\mu, f(X)). \tag{5}$$

This score, the SSW distance between the predicted and empirical measure (3), compares the spatial distributions of mass without requiring equal sample sizes or a point-to-point correspondence. For $\alpha \in (0, 1)$, the conformal quantile for the calibration scores $S_i = S(\widehat{P}_i, f(X_i))$ is defined as $\tau_\alpha = S_{(\lceil (n+1)(1-\alpha) \rceil)}$, where $S_{(k)}$ is the kth smallest calibration score and $S_{(n+1)} = \infty$. The conformal quantile τ_α immediately implies an ambient conformal prediction region

$$C_\alpha^{\mathrm{amb}}(X_{n+1}) = \big\{\mu \in P_2(\mathbb{S}^2) : S(\mu, f(X_{n+1})) \le \tau_\alpha\big\}. \tag{6}$$

Conditional on the fitted model and slices used to compute SSW (3), the calibration and test scores are exchangeable, so by the standard split conformal rank argument (Vovk et al., 2005) we have that

$$\mathbb{P}\big(\widehat{P}_{n+1} \in C_\alpha^{\mathrm{amb}}(X_{n+1})\big) \ge 1 - \alpha. \tag{7}$$

Thus, the probability that any future *empirical* measure $\widehat{P}_{n+1}$ is inside $C_\alpha^{\mathrm{amb}}(X_{n+1})$ is at least $1 - \alpha$. In the following section, we extend this guarantee to the true underlying measure P_{n+1}.

3.2 Manifold Restriction

The ambient region $C_\alpha^{\mathrm{amb}}(X_{n+1})$ is an SSW ball in $P_2(\mathbb{S}^2)$, and thus, it can admit distributions with mass between observed modes or outside the geographic constraints of the process. For example, with tropical cyclones, all predictions need to lie within specific climatological basins, while earthquakes necessarily can only occur along fault lines. Thus, in some sense, although $C_\alpha^{\mathrm{amb}}(X_{n+1})$ is statistically valid it is too large to be of practical use since it may contain vast empty regions where no data could plausibly occur.

We denote the geographic support of a distribution as a subset of $\mathbb{S}^2$ and the family of spatial measures concentrated on these structures as a subset $\mathcal{M} \subset P_2(\mathbb{S}^2)$. We refer to this family of spatial measures as the data manifold. Ideally, we would restrict $C_\alpha^{\mathrm{amb}}(X_{n+1})$ exactly to $\mathcal{M}$ to remove implausible distributions without sacrificing coverage. However, in practice, $\mathcal{M}$ is unknown and cannot be directly estimated. In fact, each training distribution is itself only observed through a small sampling of events. Instead, we can approximate proximity to $\mathcal{M}$ using the training empirical measures $\mathcal{D}_0 = \{\widehat{P}_i^0\}_{i=1}^{n_0}$. For any candidate measure μ, define the manifold discrepancy (MD) as

$$d_{\mathcal{M}}(\mu) = \min_{1 \le i \le n_0} \mathrm{SSW}_2(\mu, \widehat{P}_i^0). \tag{8}$$

Manifold discrepancy is the SSW distance to the nearest training empirical measure and provides a proxy for distance to the family of observed spatial distributions. A small value means that the candidate μ must resemble some observed distribution in its allocation of mass. Given any tolerance $\varepsilon \ge 0$, we can restrict $C_\alpha^{\mathrm{amb}}(X_{n+1})$ into a *manifold constrained* prediction set

$$C_\alpha(X_{n+1}) = \big\{\mu \in P_2(\mathbb{S}^2) : S(\mu, f(X_{n+1})) \le \tau_\alpha,\ d_{\mathcal{M}}(\mu) \le \varepsilon\big\}. \tag{9}$$

We can select a reasonable ϵ value by choosing ϵ such that it contains all training distributions, since, by definition, all training measures must be on $\mathcal{M}$. Specifically, for $n_0 \ge 2$, we select the maximum leave-one-out distance among the training empirical measures,

$$\widehat{\varepsilon} = \max_{1 \le i \le n_0} \min_{\substack{1 \le j \le n_0 \\ j \ne i}} \mathrm{SSW}_2(\widehat{P}_i^0, \widehat{P}_j^0), \tag{10}$$

and set $\varepsilon = \widehat{\varepsilon}$. This is the smallest tolerance for which every training measure has another training measure within that distance. No additional data split is required. The nearest neighbor conformal argument of Feldman et al. (2025), combined with ambient coverage, gives the following bound.

Proposition 3.1 (Manifold Coverage Lower Bound)**.** *Assume the training, calibration, and test observations are i.i.d., with $n_0 \ge 2$. Let f be fitted only to the training pairs and use slicing frames drawn independently of the data and fixed throughout. For the training leave-one-out maximum tolerance $\widehat{\varepsilon}$, take $\varepsilon = \widehat{\varepsilon}$ in equation* (9)*. Then*

$$\mathbb{P}\big(\widehat{P}_{n+1} \in C_\alpha(X_{n+1})\big) \ge 1 - \alpha - \frac{1}{n_0 + 1}.$$

The proof is in Appendix B. The empirical manifold constraint excludes the test empirical measure with probability at most $1/(n_0+1)$. The bound is marginal over the training, calibration, and test observations, and n_0 empirical measure observations. It also holds for any $\varepsilon \geq \widehat{\varepsilon}$ by set inclusion. Smaller tolerances need not satisfy this guarantee. We evaluate the choice $\varepsilon = \widehat{\varepsilon}$ in Section 4.1.

To cover the underlying distribution P_{n+1}, we allow for its discrepancy from the observed empirical measure. For a sampling error radius $r \geq 0$, define

$$C_\alpha^{(r)}(X_{n+1}) = \big\{\mu \in P_2(\mathbb{S}^2) : S(\mu, f(X_{n+1})) \leq \tau_\alpha + r,\ d_{\mathcal{M}}(\mu) \leq \widehat{\varepsilon} + r\big\}. \tag{11}$$

Both thresholds will necessarily increase by r because sampling error affects both distances.

Corollary 3.2 (Coverage of the Underlying Distribution)**.** *Assume the hypotheses of Proposition 3.1. If* $\mathbb{P}\{\mathrm{SSW}_2(P_{n+1}, \widehat{P}_{n+1}) > r\} \leq \delta$ *for* $\delta \in (0,1)$*, then*

$$\mathbb{P}\big(P_{n+1} \in C_\alpha^{(r)}(X_{n+1})\big) \geq 1 - \alpha - \frac{1}{n_0+1} - \delta.$$

The proof in Appendix C uses the triangle inequality and a union bound. Under the conditional i.i.d. location model in Section 3.1, suppose $m_{n+1} \geq m_\star \geq 1$ almost surely for a fixed lower bound $m_\star$. One valid choice is $r = b(\log(2L/\delta)/(2m_\star))^{1/4}$, where b is the circumference of the projected circle in the units used for its transport cost. This follows from the Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990) (Appendix D).

3.3 Boundary Construction by Constrained Flow

Harris (2026) introduces a flow based construction for sampling the boundary of conformal prediction regions. We extend this framework to incorporate the manifold constraint $d_{\mathcal{M}}(Y) \leq \varepsilon$, yielding a flow that targets the score boundary within the empirical manifold restricted region $C_\alpha(X_{n+1})$ defined in equation (9).

Let $\mathcal{Y}$ denote the space of finite nonempty unordered densities on $\mathbb{S}^2$. Each $Y \in \mathcal{Y}$ represent the candidate measure $\mu_Y = |Y|^{-1}\sum_{y\in Y}\delta_y$, and we abbreviate $S(Y, f(X)) = S(\mu_Y, f(X))$ and $d_{\mathcal{M}}(Y) = d_{\mathcal{M}}(\mu_Y)$. Let $Y(t) \in \mathcal{Y}$ be a time indexed density and let $y(t) \in \mathbb{R}^{3|Y|}$ denote the vectorization of its point coordinates. Define the two scalar quantities

$$S(y) = S(Y, f(X_{n+1})) = \mathrm{SSW}_2(\mu_Y, f(X_{n+1})), \qquad d(y) = d_{\mathcal{M}}(Y), \tag{12}$$

with gradients $g_S(y) = \nabla_y S$ and $g_d(y) = \nabla_y d$ obtained by automatic differentiation through the SSW computation. Each point in the candidate cloud is matched to its geodesic nearest neighbor in the pooled empirical reference support, and the manifold discrepancy is evaluated by SSW between the candidate cloud and this locally matched reference cloud. We seek a velocity field $v(y) = \dot{y}$ such that the trajectory $Y(t)$ converges to the set $\{S = \tau_\alpha,\ d \leq \varepsilon\}$. We specify this by imposing target scalar dynamics on both quantities:

$$\frac{d}{dt}S(Y(t)) = -\lambda_1\big(S(Y(t)) - \tau_\alpha\big), \tag{13}$$

$$\frac{d}{dt}d_{\mathcal{M}}(Y(t)) = -\lambda_2\big(d_{\mathcal{M}}(Y(t)) - \varepsilon\big)_+, \tag{14}$$

where $\lambda_1, \lambda_2 > 0$ are decay rates and $(u)_+ = \max\{u, 0\}$. The first equation drives S exponentially toward τ_α from any initial value; at steady state $S(Y) = \tau_\alpha$, placing the density on the conformal boundary. The second equation drives d toward ε only when $d > \varepsilon$, acting as a one sided restoring force that enforces the manifold tube while remaining tangent to the level set shell. When both scalar dynamics are attained and converge, μ_Y lies on the conformal shell within the manifold tube. Due to the empirical sliced Wasserstein objective being piecewise differentiable, we additionally interpret the gradient in the velocity field construction as a subgradient at nondifferentiable configurations.

By the chain rule, equations (13)–(14) impose two linear constraints on $v(y)$:

$$J(y)\,v(y) = -c(y), \qquad J(y) = \begin{pmatrix} g_S(y)^\top \\ g_d(y)^\top \end{pmatrix} \in \mathbb{R}^{2\times 3|Y|}, \quad c(y) = \begin{pmatrix} \lambda_1\big(S(Y) - \tau_\alpha\big) \\ \lambda_2\big(d_{\mathcal{M}}(Y) - \varepsilon\big)_+ \end{pmatrix}. \tag{15}$$

Since J has two rows and $3|Y|$ columns, a consistent system $Jv = -c$ has infinitely many solutions. We take the minimum norm solution, given by the Moore–Penrose pseudoinverse:

$$v(y) = -J(y)^\top \{J(y)J(y)^\top\}^+ c(y). \tag{16}$$

Derivation in Appendix A. For a consistent system, this choice satisfies the prescribed scalar dynamics without motion orthogonal to both g_S and g_d. When the constraints are inconsistent, it instead gives the minimum norm least squares velocity. The flow is integrated by forward Euler steps as Harris (2026), initialized at a realization sampled from the forecasting model's prediction for the test observation. After each step, every point in Y is projected back onto $\mathbb{S}^2$ by normalization to ensure the trajectory remains on the sphere. We retain the manifold constraint throughout the flow, including when the current iterate lies inside the tube. This conservative choice preserves feasibility while avoiding switching behavior near the constraint boundary.

The velocity field in equation (16) represents candidate measures at a fixed conformal level whenever the score boundary intersects the empirical tube. Following Harris (2026), we combine the output of the flow across levels to form an ensemble over spatial measures,

$$P_x^{\mathrm{CPD}}(A) = \int_0^1 \nu_{x,\alpha}(A)\, d\alpha, \qquad A \in \mathcal{B}(P_2(\mathbb{S}^2)), \tag{17}$$

where $\nu_{x,\alpha}$ is the distribution of the empirical measure obtained from the numerical flow at level α. The base distribution assigns equal mass to the calibration densities, and is initialized at the test forecast density. Thus each draw from P_x^{CPD} is a spatial measure represented by a generating empirical density. The coverage results apply to the sets in equations (9) and (11); they do not establish calibration of arbitrary summaries of this ensemble. To target the enlarged set, the same velocity formula uses thresholds $\tau_\alpha + r$ and $\widehat{\varepsilon} + r$.

4 Numerical Studies

We evaluate empirical measure coverage (Proposition 3.1) on synthetic processes and compare the ensembles in equation (17) on synthetic and real geophysical data. The reported experiments use the empirical region C_α, without the sampling error enlargement in Corollary 3.2. The tropical cyclone and earthquake studies use chronological splits, so their results assess predictive performance without establishing the i.i.d. coverage guarantee.

Synthetic event samples are generated from four spatial processes: uniform noise over the sphere (UWN), a misspecified spiral distribution (Spiral), a three mode density (3-Mode), and a Hawkes process. These processes represent increasing levels of spatial heterogeneity and dependence. Each event sample is represented by its normalized empirical measure. For the ensemble comparisons, we generate 50 observations in each of the training, calibration, and test sets and use a simple 2D histogram model for prediction. Full simulation settings are described in Appendix F.

4.1 Empirical Coverage Lower Bound

We first evaluate the empirical coverage of our procedure under the four simulation settings: UWN, Spiral, 3-Mode, and the Hawkes under a range of hyperparameter settings. Figure 2 shows coverage of the observed empirical measures as we vary the number of SSW slices L, events per observation n, calibration observations n_{cal}, and manifold tolerance ε. Here the figure notation n corresponds to m_i in equation (4). If ε is set too low, then our approach can severely under cover the target; Proposition 3.1 guarantees coverage of at least $1 - \alpha - 1/(n_0 + 1)$ only for $\varepsilon \geq \widehat{\varepsilon}$. The number of slices L has little to no effect, but should be set high enough to ensure stable estimation of SSW.

4.2 Synthetic Data Evaluations

We use the four data generating mechanisms in Section 4.1, UWN, Spiral, 3-Mode, and Hawkes, to evaluate the predictive distribution produced by an unconstrained velocity field (VF) (Harris, 2026) and a manifold constrained velocity field (CVF) (equation (16)). For each process, we generate 50 training, calibration, and test observations, each containing 50 or 100 events. For our predictive model, we construct a simple predictor by gridding the training observations into cells and normalizing the cell counts into probabilities, i.e. a 2D histogram (Appendix E).

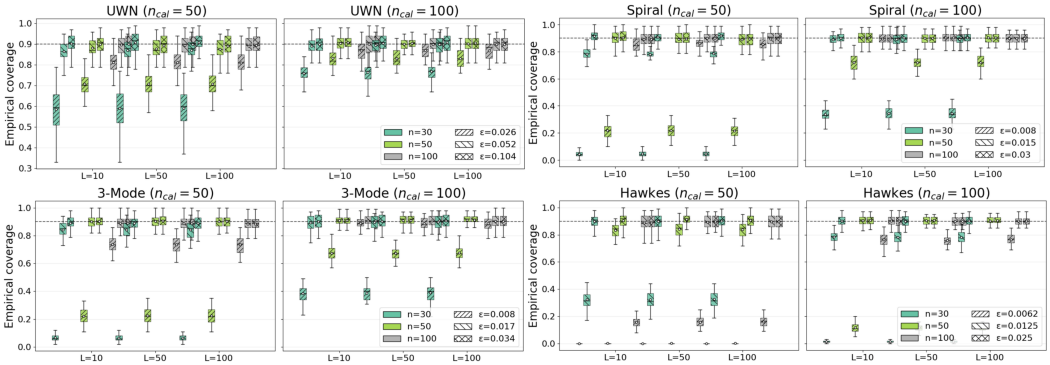

Figure 2: Coverage by calibration size (n_{cal}), empirical sample size (n), SSW slices (L), and manifold tolerance (100 slices ε) over test simulations including the leave one out maximum manifold value (Middle ε value).

We compare VF and CVF (Section 3.3) against five baseline approaches. These include a minimum area highest density region (HDR) estimator inspired by classical HDR and bayesian HPD sets (M1)(Hyndman, 1996; Polonik, 1997), a conformalized minimum area HDR using APS (M2) (Romano et al., 2020), and three standard generative models: Diffusion (M3) (Ho et al., 2020), Flow Matching (M4) (Lipman et al., 2023; Chen & Lipman, 2024), and a Variational Auto Encoder (VAE) (M5) (Kingma & Welling, 2013). The two HDR baselines realizations are obtained by sampling from the forecast distribution restricted to the selected region, followed by uniform cell jitter before conversion to spherical coordinates. VAE initialized from the forecast representation and trained on matched forecast/observation pairs. (Appendix E).

Table 1: Simulation results across four settings. MinArea HDR (M1), Conformal MinArea HDR (M2), Diffusion (M3), Flow Matching (M4), VAE (M5), velocity field (VF), and constrained velocity field (CVF). Metrics are: Energy distance (ED) and Manifold distance (MD). Standard errors are reported in parentheses.

Process	Metric	M1	M2	M3	M4	M5	VF	CVF
Spiral	**ED** $\times 10^2$	6.69 (2.21)	6.40 (2.22)	6.86 (2.24)	3.85 (1.71)	6.24 (2.21)	0.57 (0.36)	0.55 (0.41)
	Pred–Pred	3.98 (-)	4.13 (-)	4.10 (-)	10.85 (-)	4.29 (-)	23.99 (-)	23.60 (-)
	True–Pred	15.98 (1.44)	15.91 (1.44)	16.13 (1.45)	18.00 (1.18)	15.91 (1.44)	22.93 (0.50)	22.72 (0.53)
	MD $\times 10^2$	1.29 (0.01)	1.27 (0.01)	1.16 (0.01)	1.27 (0.02)	1.19 (0.01)	1.52 (0.09)	1.50 (0.06)
	ED $\times 10^2$	0.86 (0.12)	0.68 (0.12)	0.90 (0.17)	0.20 (0.09)	0.74 (0.15)	$8.2 * 10^{-5}$ (0.09)	$7.7 * 10^{-5}$ (0.09)
	Pred–Pred	4.966 (-)	4.893 (-)	5.335 (-)	5.248 (-)	5.017 (-)	5.498 (-)	5.492 (-)
	True–Pred	5.56 (0.09)	5.43 (0.10)	5.76 (0.11)	5.37 (0.08)	5.52 (0.11)	5.40 (0.08)	5.40 (0.08)
	MD $\times 10^2$	1.90 (0.032)	1.86 (0.029)	1.61 (0.018)	1.88 (0.023)	1.86 (0.025)	1.00 (0.049)	0.94 (0.044)
3-Mode	**ED** $\times 10^3$	-0.42 (1.31)	-0.23 (1.51)	3.29 (1.00)	15.51 (1.39)	-0.34 (1.27)	0.58 (1.02)	0.55 (1.02)
	Pred–Pred	60.63 (-)	54.73 (-)	60.27 (-)	55.86 (-)	56.93 (-)	70.68 (-)	70.59 (-)
	True–Pred	58.35 (1.32)	55.49 (1.42)	60.02 (1.11)	63.92 (1.15)	56.54 (1.31)	63.87 (1.15)	63.81 (1.16)
	MD $\times 10^2$	0.95 (0.034)	1.02 (0.024)	1.19 (0.035)	1.61 (0.045)	1.05 (0.024)	0.91 (0.035)	0.85 (0.034)
Hawkes	**ED** $\times 10^3$	2.41 (1.76)	1.51 (1.68)	3.57 (1.96)	3.48 (2.05)	1.34 (1.64)	-0.26 (1.33)	-0.24 (1.33)
	Pred–Pred	41.67 (-)	42.36 (-)	42.33 (-)	42.94 (-)	39.92 (-)	54.68 (-)	54.65 (-)
	True–Pred	49.96 (1.40)	49.86 (1.36)	50.88 (1.43)	51.14 (1.46)	48.56 (1.36)	55.14 (1.18)	55.13 (1.18)
	MD $\times 10^2$	0.85 (0.01)	0.86 (0.01)	0.88 (0.01)	0.88 (0.01)	0.86 (0.01)	0.80 (0.02)	0.80 (0.02)

We evaluate each ensemble by an energy distance style statistic (Rizzo & Székely, 2016; Sejdinovic et al., 2013) comparing it with the test empirical measures, and by its distance to the estimated data manifold. We use the spherical sliced Wasserstein in place of the Euclidean ground distance,

$$\mathrm{ED}(\mu, \nu) = 2\mathbb{E}[\mathrm{SSW}_2(\mu, \nu)] - \mathbb{E}[\mathrm{SSW}_2(\nu, \nu')] - \mathbb{E}[\mathrm{SSW}_2(\mu, \mu')], \tag{18}$$

where μ, μ' are independent draws from the distribution of test empirical measures and ν, ν' are independent draws from the predictive ensemble, with both pairs independent. For generated samples, we report the local numerical manifold discrepancy used during flow integration. We provide a Euclidean energy score on consistently gridded probability vectors in Appendix G.1 for robustness.

Table 1 shows the average ED and MD over the test data. CVF attains the lowest MD in the Misspecified Spiral setting and has all around great performance along with VF. The manifold constraint has little effect in the uniform setting. In the structured settings, controlling distance

to the training measures preserves ED relative to VF while reducing MD. The two criteria assess distributional agreement and proximity to the observed family of spatial distributions, respectively.

4.3 Tropical Cyclones

We now apply our method to the spatial distribution of tropical cyclone genesis in each season (Qian et al., 2022). Each season provides one empirical measure of genesis locations, and the predictive ensemble represents variation among seasonal spatial distributions. Tropical cyclone events are obtained from the IBTrACS database (Gahtan et al., 2026), which records the time and location where each storm attains tropical cyclone strength and status (Qian et al., 2022).

We train a simple prediction model (Appendix G.6) on the 1980 to 1989 data to generate yearly global probabilistic forecast maps. We then use the 1990 to 2010 data as calibration for each conformal method (Sec. 4.2) and fit each of the baseline methods, VF, and CVF. For CVF we set $\varepsilon = \widehat{\varepsilon}$ from equation (10). Over the 2011-2023 test set, 11 of the 13 observed seasonal distributions fall within the conformal SSW ball defined by the calibration data. In addition, we uniformly sample α from 1 to 0.1 for level set creation, and hold $\alpha = 0.1$ for our method and HDR based methods.

Table 2: Tropical cyclone energy distance metrics

Metric	M1	M2	M3	M4	M5	VF	CVF
ED	3.89	4.10	0.88	1.62	0.63	0.42	0.53
Pred–Pred	4.91	4.04	5.60	5.55	6.82	4.82	4.73
True–Pred	6.94	6.61	5.78	6.12	6.26	5.16	5.17
MD	1.45	1.49	1.79	2.34	1.81	1.65	0.93

Table 2 shows the average energy distance (ED) (Equation 18), broken down by its constituent terms, and the average manifold distance (MD) over the test period (2011 - 2023). The CVF and VF methods achieve the best ED scores compared to any of the baseline or generative methods. However, CVF also has the lowest MD (0.93), compared with (1.65) for VF, indicating closer agreement with the family of observed seasonal distributions.

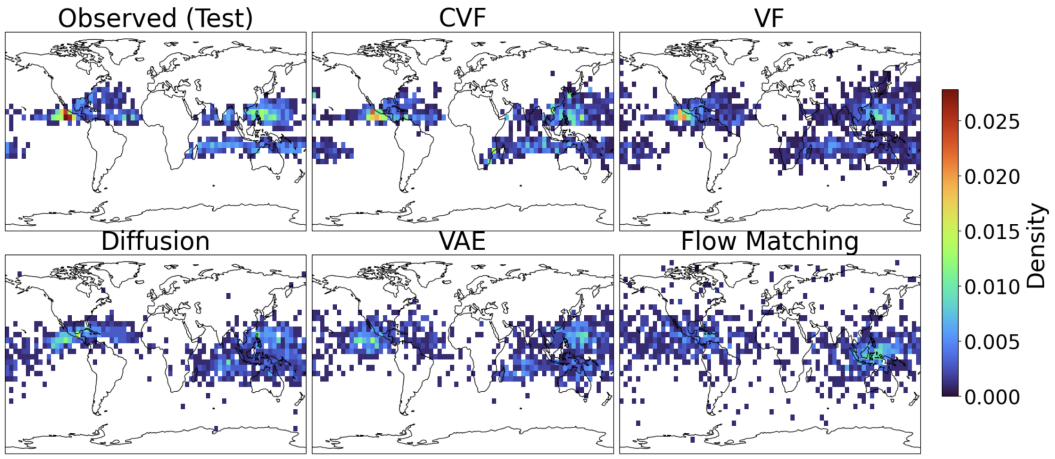

Figure 3: 2023 prediction regions of 5 methods with all observed tropical cyclones from 2011-2023

Figure 3 shows an example density from each method (Year 2023). The CVF density (3b) is highly concentrated in the observed tropical cyclone basins, while VF is too dispersed with support far beyond the observed tropical cyclones (3a). The other sampling methods all obviously over support tropical cyclone genesis locations.

4.4 Earthquake Events

Finally, we apply our method to earthquake events in California. Our data comes from the ANSS Comprehensive Earthquake Catalog (U.S. Geological Survey, 2026), which we restrict to only magnitude ≥ 4.0 events. Each year provides one empirical measure of earthquake locations. Earthquake activity

concentrates probability mass along fault systems, so uncertainty in the annual spatial distribution should preserve this structure.

As in Section 4.3 we use a cumulative 2D empirical histogram predictor based on the forecasting model(Werner et al., 2010) up to the year 1985 where we start calibration in G.7. We compare VF and CVF against baselines in Section 4.3 again using ED and MD along with $\alpha = 0.1$ and region construction. We train the prediction model on 1976-1984 data, calibrate on 1985-2010 data, and test on 2011-2025 data where all 15 members of the test set fall within the conformal $\alpha = 0.1$ SSW ball.

Table 3: California earthquake energy distance metrics

Metric	M1	M2	M3	M4	M5	VF	CVF
ED $\times 100$	0.32	0.32	0.76	0.50	0.54	0.12	0.15
Pred–Pred	1.04	1.07	1.29	1.07	1.10	1.83	1.83
True–Pred	1.52	1.54	1.87	1.63	1.67	1.82	1.83
MD $\times 100$	0.53	0.68	1.38	0.99	1.12	0.78	0.54

Table 3 shows average ED and MD over the test period. VF and CVF have the lowest ED (0.12 and 0.15), while M1 and CVF have the lowest MD (0.53 and 0.54). Relative to VF, CVF reduces MD from 0.78 to 0.54 with only a small increase in ED.

4.5 Sample Diversity Within Conformal Level Set

In addition to level coverage, the proposed conformal sampling method captures large variation in the spatial structure of plausible densities. At the 90% level, we find admissible tropical cyclone genesis distributions that place a majority of their mass in completely different geographic basins, which yield opposite conclusions yet lie on the same level set. In this setting, this sample diversity is desirable due to the inclusion of qualitatively distinct spatial structures in addition to providing mass in different geographic regions, even of which, a specific forecast density overlooks. This is further corroborated in Appendix G.8 by comparing variances of SSW of level sets to each density quantile.

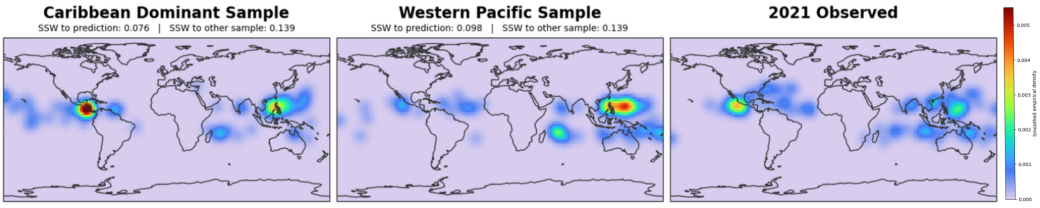

Figure 4: 2021 example of sample diversity at $\alpha = 0.1$ level set

5 Discussion

We constructed conformal prediction sets for spatial distributions observed through finite event samples. The sets restrict an SSW ball around the forecast to a union of balls around the training measures, at a coverage cost of at most $1/(n_0+1)$ and with no additional data split, and a constrained flow represents them as ensembles of spatial densities on a sphere. For tropical cyclone genesis locations and California earthquakes, the restriction concentrates the ensemble on the observed basins and fault systems while keeping energy distance at or near that of the unconstrained flow.

There are a few limitations to our current work. The real data case studies use chronological splits, which may not satisfy the exchangeability assumption of Proposition 3.1. However, the empirical performance of the sampled ensembles is still quite competitive. The sampling error radius of Appendix D is uninformative at the event counts studied here, and tighter radii, including radii for dependent events, are needed before the guarantee for the underlying distribution is usable in practice. Finally, the training empirical measures are a coarse surrogate for the family of plausible distributions, and better estimates of that family could shrink the set further.

6 AI Disclosure Statement

In this work, we used generative AI tools to assist with the design and implementation of synthetic experiments and corroborating mathematical claims. We did not use generative AI to generate the primary research question, main methodological contribution, or independently establish the theoretical results presented in this paper. We additionally used generative AI tools to polish and improve on the readability of portions of the manuscript and to assist in discovery of related literature. Each source of literature was manually checked against the original source prior to inclusion. In addition all mathematical statements and AI assisted experimentation code was reviewed and verified by the authors. The authors made the final decisions regarding design, mathematical arguments, results interpretation and content. We take responsibility for the final content of this work including all text, code, and experiments assisted with generative AI.

References

Peyman Abbaszadeh, David F Munoz, Hamed Moftakhari, Keighobad Jafarzadegan, and Hamid Moradkhani. Perspective on uncertainty quantification and reduction in compound flood modeling and forecasting. *iScience*, 25(10), 2022.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer, 2005.

Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. Interdisciplinary Statistics. Chapman and Hall/CRC, Boca Raton, FL, 2015.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Keith J Beven, Willy P Aspinall, Paul D Bates, Edoardo Borgomeo, Katsuichiro Goda, Jim W Hall, Trevor Page, Jeremy C Phillips, Michael Simpson, Paul J Smith, et al. Epistemic uncertainties and natural hazard risk assessment—Part 2: What should constitute good practice? *Natural Hazards and Earth System Sciences*, 18(10):2769–2783, 2018.

Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh-Tan Pham. Spherical sliced-wasserstein. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jXQ0ipgMdU.

George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1973.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024. OpenReview ID: g7ohDlTITL.

Chicago Police Department. Crimes – 2001 to present. City of Chicago Data Portal, 2026. Dataset ID: ijzp-q8t2. Accessed September 24, 2026.

Anthony C Davison and Raphaël Huser. Statistics of extremes. *Annual Review of Statistics and its Application*, 2(1):203–235, 2015.

Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

Aaron O. Feldman, Joseph A. Vincent, Maximilian Adang, Jun En Low, and Mac Schwager. Learning robot safety from sparse human feedback using conformal prediction. *arXiv preprint arXiv:2501.04823*, 2025.

Jennifer Gahtan, Kenneth R. Knapp, Carl J. III Schreck, Howard J. Diamond, James P. Kossin, and Michael C. Kruk. International best track archive for climate stewardship (IBTrACS) project, version 4.01, 2026. Dataset. DOI: 10.25921/82ty-9e16. Available at `https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc%3AC01552`. Accessed 2026-02-23.

Trevor Harris. Flow-based conformal predictive distributions. *arXiv preprint arXiv:2602.07633*, 2026.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58 (1):83–90, 1971.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2): 120–126, 1996. doi: 10.1080/00031305.1996.10474359.

Thomas H Jordan, Yun-Tai Chen, Paolo Gasparini, Raul Madariaga, Ian Main, Warner Marzocchi, Gerassimos Papadopoulos, Gennady Sobolev, Koshun Yamaoka, and Jochen Zschau. Operational earthquake forecasting: State of knowledge and guidelines for utilization. *Annals of Geophysics*, 54(4):315–391, 2011.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Michal Klein, Louis Bethune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport. *arXiv preprint arXiv:2502.03609*, 2025.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1):29–43, 2015.

Martin Leutbecher and Tim N Palmer. Ensemble forecasting. *Journal of Computational Physics*, 227 (7):3515–3539, 2008.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Gabriel Melo, Thibaut de Saivre, Anna Calissano, and Florence d'Alché-Buc. Conformal graph prediction with Z-Gromov–Wasserstein distances. *arXiv preprint arXiv:2603.02460*, 2026.

Bruno Merz, Christian Kuhlicke, Michael Kunz, Massimiliano Pittore, Andrey Babeyko, David N Bresch, Daniela IV Domeisen, Frauke Feser, Inga Koszalka, Heidi Kreibich, et al. Impact forecasting to support emergency management of natural hazards. *Reviews of Geophysics*, 58(4): e2020RG000704, 2020.

Sumantrak Mukherjee, Mouad Elhamdi, George Mohler, David A Selby, Yao Xie, Sebastian Vollmer, and Gerrit Grossmann. Neural spatiotemporal point processes: Trends and challenges. *arXiv preprint arXiv:2502.09341*, 2025.

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

Muhammad Osama, Dave Zachariah, and Peter Stoica. Prediction of spatial point processes: Regularized method with out-of-sample guarantees. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11919–11928, 2019.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Now Foundations and Trends, 2019.

Wolfgang Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1–24, July 1997. doi: 10.1016/S0304-4149(97)00028-8.

QiFeng Qian, XiaoJing Jia, and Yanluan Lin. Reduced tropical cyclone genesis in the future as predicted by a machine learning model. *Earth's Future*, 10(2):e2021EF002455, 2022. doi: https://doi.org/10.1029/2021EF002455. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021EF002455. e2021EF002455 2021EF002455.

Maria L. Rizzo and Gábor J. Székely. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38, 2016. doi: 10.1002/wics.1375.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591. Curran Associates, Inc., 2020.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pp. 2263–2291, 2013.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Max Sommerfeld and Axel Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):219–238, 2018. doi: 10.1111/rssb.12236.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. *arXiv preprint arXiv:2501.18991*, 2025.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

U.S. Geological Survey. ANSS comprehensive earthquake catalog (ComCat). https://earthquake.usgs.gov/data/comcat/, 2026. Accessed 2026-04-24.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019. doi: 10.3150/18-BEJ1065.

Maximilian J. Werner, Agnes Helmstetter, David D. Jackson, Yan Y. Kagan, and Stefan Wiemer. Adaptively smoothed seismicity earthquake forecasts for italy. *Annals of Geophysics*, 53(3):107–116, November 2010. ISSN 1593-5213. doi: 10.4401/ag-4839. URL http://dx.doi.org/10.4401/ag-4839.

Rui Xu, Chao Chen, Yue Sun, Parvathinathan Venkitasubramaniam, and Sihong Xie. Wasserstein-regularized conformal prediction under general distribution shift. In *International Conference on Learning Representations*, 2025.

Zihao Zhou, Xingyi Yang, Ryan Rossi, Handong Zhao, and Rose Yu. Neural point process for learning spatiotemporal event dynamics. In *Learning for Dynamics and Control Conference*, pp. 777–789, 2022.

Jakob Zscheischler, Seth Westra, Bart JJM Van Den Hurk, Sonia I Seneviratne, Philip J Ward, Andy Pitman, Amir AghaKouchak, David N Bresch, Michael Leonard, Thomas Wahl, et al. Future climate risk from compound events. *Nature Climate Change*, 8(6):469–477, 2018.

A Derivation of the Velocity Field

We derive the minimum norm velocity field used in Section 3.3. Fix X_{n+1} and write

$$S(y) = S(Y, f(X_{n+1})), \qquad d(y) = d_M(Y),$$

where $y \in \mathbb{R}^{3|Y|}$ is the vectorized point cloud. The desired scalar dynamics are

$$\frac{d}{dt}S(y(t)) = -\lambda_1\{S(y(t)) - \tau_\alpha\}, \qquad \frac{d}{dt}d(y(t)) = -\lambda_2\{d(y(t)) - \varepsilon\}_+.$$

If $v(y) = y'(t)$, then the chain rule gives

$$\frac{d}{dt}S(y(t)) = \nabla S(y)^\top v(y), \qquad \frac{d}{dt}d(y(t)) = \nabla d(y)^\top v(y).$$

Thus the velocity must satisfy the two linear constraints

$$\nabla S(y)^\top v(y) = -\lambda_1\{S(y) - \tau_\alpha\}, \qquad \nabla d(y)^\top v(y) = -\lambda_2\{d(y) - \varepsilon\}_+.$$

Equivalently, with

$$J(y) = \begin{pmatrix} \nabla S(y)^\top \\ \nabla d(y)^\top \end{pmatrix}, \qquad c(y) = \begin{pmatrix} \lambda_1\{S(y) - \tau_\alpha\} \\ \lambda_2\{d(y) - \varepsilon\}_+ \end{pmatrix},$$

these constraints become $J(y)v(y) = -c(y)$.

Since $J(y) \in \mathbb{R}^{2\times 3|Y|}$, the system is generally underdetermined. We select the minimum norm solution,

$$v^\star(y) = \arg\min_{v\in\mathbb{R}^{3|Y|}} \frac{1}{2}\|v\|_2^2 \quad \text{subject to} \quad J(y)v = -c(y).$$

The Lagrangian is $\mathcal{L}(v, \xi) = \frac{1}{2}v^\top v + \xi^\top\{J(y)v + c(y)\}$, where $\xi \in \mathbb{R}^2$ is the vector of Lagrange multipliers. Stationarity with respect to v gives $\partial\mathcal{L}/\partial v = v + J(y)^\top\xi = 0$, hence $v = -J(y)^\top\xi$. Substituting back into the constraint $J(y)v = -c(y)$ gives

$$-J(y)J(y)^\top\xi = -c(y), \qquad \text{so} \qquad J(y)J(y)^\top\xi = c(y).$$

When $J(y)$ has full row rank (i.e. $\nabla_y S$ and $\nabla_y d$ are linearly independent), $J(y)J(y)^\top \in \mathbb{R}^{2\times 2}$ is invertible, so $\xi = \{J(y)J(y)^\top\}^{-1}c(y)$, and therefore

$$v^\star(y) = -J(y)^\top\{J(y)J(y)^\top\}^{-1}c(y).$$

In the rank deficient case, $\nabla_y S$ and $\nabla_y d$ are parallel, so $J(y)J(y)^\top$ is singular. The constraint $J(y)v = -c(y)$ is then either inconsistent (if $c(y)$ is not in the column space of $J(y)J(y)^\top$) or has infinitely many solutions. Among the minimum norm solutions in the consistent case, we replace the inverse with the Moore–Penrose pseudoinverse:

$$v^\star(y) = -J(y)^\top\{J(y)J(y)^\top\}^{+}c(y).$$

This recovers the full rank solution when $J(y)$ has full row rank, and otherwise projects $c(y)$ onto the column space of $J(y)J(y)^\top$ before solving, giving the minimum norm least squares velocity.

The full rank expression satisfies the prescribed scalar dynamics exactly, since

$$J(y)v^\star(y) = -J(y)J(y)^\top\{J(y)J(y)^\top\}^{-1}c(y) = -c(y).$$

Consequently,

$$\frac{d}{dt}S(y(t)) = -\lambda_1\{S(y(t)) - \tau_\alpha\}, \qquad \frac{d}{dt}d(y(t)) = -\lambda_2\{d(y(t)) - \varepsilon\}_+.$$

For numerical stability we use the damped form

$$v_\eta(y) = -J(y)^\top\{J(y)J(y)^\top + \eta I\}^{-1}c(y),$$

with $\eta > 0$. This is the regularized version of the same minimum norm controller and converges to the exact expression as $\eta \downarrow 0$ whenever $J(y)J(y)^\top$ is nonsingular.

B Proof of Proposition 3.1

Proof. Write $N = n_0 + 1$, $Z_i = Y_i^0$ for $i < N$, and $Z_N = Y_{n+1}$. Using the fixed SSW slicing frames, define the nearest neighbor distances in the augmented sample by

$$R_i = \min_{\substack{1 \le j \le N \\ j \ne i}} \mathrm{SSW}_2(\mu_{Z_i}, \mu_{Z_j}), \qquad i = 1, \dots, N.$$

The vector $(R_1, \dots, R_N)$ is exchangeable. For $i < N$, adding the test cloud can only decrease the nearest neighbor distance. Since $\widehat{\varepsilon}$ is the maximum training leave-one-out distance,

$$R_i \le \min_{\substack{1 \le j < N \\ j \ne i}} \mathrm{SSW}_2(\mu_{Z_i}, \mu_{Z_j}) \le \widehat{\varepsilon}, \qquad i < N.$$

Also, $R_N = d_{\mathcal{M}}(\widehat{P}_{n+1})$. Thus $E = \{d_{\mathcal{M}}(\widehat{P}_{n+1}) > \widehat{\varepsilon}\}$ implies that R_N is the unique maximum. By exchangeability,

$$\begin{aligned} \mathbb{P}(E) &\le \mathbb{P}\Big(R_N > \max_{i<N} R_i\Big) \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}\Big(R_i > \max_{j \ne i} R_j\Big) \le \frac{1}{N}, \end{aligned}$$

because at most one index can attain a strict maximum. Ties do not affect this bound. This is the maximum threshold case of the nearest neighbor conformal result of Feldman et al. (2025, Theorem V.1).

With $\varepsilon = \widehat{\varepsilon}$, exclusion from $C_\alpha(X_{n+1})$ requires either exclusion from the ambient region or the event E. The ambient coverage guarantee in Section 3.1 therefore gives

$$\begin{aligned} \mathbb{P}\big(\widehat{P}_{n+1} \notin C_\alpha(X_{n+1})\big) &\le \mathbb{P}\big(\widehat{P}_{n+1} \notin C_\alpha^{\mathrm{amb}}(X_{n+1})\big) + \mathbb{P}(E) \\ &\le \alpha + \frac{1}{n_0 + 1}. \end{aligned}$$

Taking complements proves the result. Independence between the two exclusion events is not required. □

C Proof of Corollary 3.2

Proof. On the event $\mathrm{SSW}_2(P_{n+1}, \widehat{P}_{n+1}) \le r$, the triangle inequality gives

$$\begin{aligned} S(P_{n+1}, f(X_{n+1})) &\le S(\widehat{P}_{n+1}, f(X_{n+1})) + r, \\ d_{\mathcal{M}}(P_{n+1}) &\le d_{\mathcal{M}}(\widehat{P}_{n+1}) + r. \end{aligned}$$

The second inequality follows by applying the triangle inequality to every training reference measure and taking the minimum. Thus $\widehat{P}_{n+1} \in C_\alpha(X_{n+1})$ and $\mathrm{SSW}_2(P_{n+1}, \widehat{P}_{n+1}) \le r$ together imply $P_{n+1} \in C_\alpha^{(r)}(X_{n+1})$. Proposition 3.1 and a union bound yield

$$\mathbb{P}\big(P_{n+1} \notin C_\alpha^{(r)}(X_{n+1})\big) \le \alpha + \frac{1}{n_0 + 1} + \delta.$$

The events need not be independent. The training references remain empirical in both sets, so this argument requires an approximation bound only for the test distribution. □

D A Sampling Error Radius for Independent Events

Suppose that, conditional on $(P_{n+1}, X_{n+1}, m_{n+1})$, the test locations are i.i.d. from P_{n+1}, and that $m_{n+1} \ge m_\star \ge 1$ almost surely. Fix the L slicing frames independently of the data. Parameterize each projected circle by an interval $[0, b)$, where b is its circumference in the units used for transport: $b = 2\pi$ for radians and $b = 1$ for angles normalized to one revolution. Let F_ℓ and $\widehat{F}_\ell$ be the distribution

functions of the projected population and empirical measures in that interval. The Dvoretzky–Kiefer–Wolfowitz inequality (Massart, 1990), conditional on the test distribution, covariate, sample size, and slices, yields

$$\mathbb{P}\left\{\max_{1\le\ell\le L}\|\widehat{F}_\ell - F_\ell\|_\infty > t\right\} \le 2L\exp(-2m_\star t^2).$$

The union bound over projections does not require independence among slices. For probability measures λ, λ' represented on $[0, b)$, geodesic distance on the circle is bounded by distance on the interval. Writing $W_{p,[0,b]}$ for interval transport and $W_{2,\mathrm{circ}}$ for circle transport, we have

$$W^2_{2,\mathrm{circ}}(\lambda,\lambda') \le W^2_{2,[0,b]}(\lambda,\lambda') \le bW_{1,[0,b]}(\lambda,\lambda') \le b^2\|F_\lambda - F_{\lambda'}\|_\infty.$$

The middle bound uses $|u-v|^2 \le b|u-v|$ in a coupling optimal for W_1, and the last uses the integral formula for one dimensional W_1. Substituting $t = \{\log(2L/\delta)/(2m_\star)\}^{1/2}$ into the probability bound and averaging the squared circle distances in equation (3) gives

$$\mathbb{P}\left\{\mathrm{SSW}_2(P_{n+1}, \widehat{P}_{n+1}) > b\left(\frac{\log(2L/\delta)}{2m_\star}\right)^{1/4}\right\} \le \delta.$$

This bound is uniform over the underlying measure. It remains valid after averaging over the test distribution, covariate, sample size, and frames. The fixed lower bound $m_\star$ makes the radius available before the test events are observed.

The $m_\star^{-1/4}$ dependence is consistent with known worst case empirical Wasserstein rates on compact metric spaces (Weed & Bach, 2019) and, in particular, with the nonstandard empirical W_2 behavior that arises for discrete or separated supports (Sommerfeld & Munk, 2018). Thus, without additional regularity assumptions on the projected distributions, a uniform $m_\star^{-1/2}$ rate for W_2 cannot generally be expected. Our bound is intended as a distribution-free sampling correction; its constants and dependence on the number of slices may nevertheless be conservative.

E Density Valued Predictions

The conformity score $S(Y, f(X)) = \mathrm{SSW}_2(\mu_Y, f(X))$ requires both the observed target and the model prediction to be representable as elements of $P_2(\mathbb{S}^2)$. For observed targets this is immediate since any finite point cloud Y maps to its empirical measure μ_Y as defined above. For predictions the situation is more varied. Many forecasting systems do not produce point cloud samples but instead output a continuous probability density $p : \mathbb{S}^2 \to \mathbb{R}_+$ over event locations, for example as a normalized spatial intensity map from a neural network, a kernel density estimate, or a gridded output from a climate model. In all such cases the prediction is already an element of $P_2(\mathbb{S}^2)$, and the SSW score applies directly once the density is discretized to be compatible with the sliced Wasserstein computation.

We discretize the predicted density onto a regular latitude longitude grid. Parameterize $\mathbb{S}^2$ by longitude $\lambda \in [-\pi, \pi)$ and latitude $\varphi \in (-\pi/2, \pi/2)$ via $x(\lambda, \varphi) = (\cos\varphi\cos\lambda,\ \cos\varphi\sin\lambda,\ \sin\varphi)$, with surface area element $d\sigma = \cos\varphi\, d\lambda\, d\varphi$. Fix grid edges $\{\lambda_i\}_{i=0}^{N_\lambda}$ and $\{\varphi_j\}_{j=0}^{N_\varphi}$ partitioning the domain, and define rectangular cells

$$B_{ij} = \{x \in \mathbb{S}^2 : \lambda_i \le \lambda(x) < \lambda_{i+1},\ \varphi_j \le \varphi(x) < \varphi_{j+1}\}. \tag{19}$$

The probability mass assigned to bin (i, j) is

$$q_{ij} = \int_{B_{ij}} p(x)\, d\sigma(x) \approx p(c_{ij})\ \cos\bar{\varphi}_j\, \Delta\lambda\, \Delta\varphi, \tag{20}$$

where $c_{ij} = x(\bar{\lambda}_i, \bar{\varphi}_j)$ is the bin center at midpoint coordinates $\bar{\lambda}_i = (\lambda_i + \lambda_{i+1})/2$, $\bar{\varphi}_j = (\varphi_j + \varphi_{j+1})/2$, and the right hand approximation holds when p is smooth relative to the grid resolution. The weights $\{q_{ij}\}$ define a discrete measure

$$\nu = \sum_{i,j} q_{ij}\, \delta_{c_{ij}} \in P_2(\mathbb{S}^2), \tag{21}$$

which is a valid element of the Wasserstein space $P_2(\mathbb{S}^2)$ and converges weakly to the continuous measure $p\,d\sigma$ as the grid is refined. The SSW distance between ν and any empirical measure μ_Y is then computed exactly as in the score definition, with the quantized measure replacing $f(X_{n+1})$.

To draw a point cloud realization from ν for use in the constrained flow, we sample bin indices $(i, j) \sim \text{Categorical}(\{q_{ij}\})$ independently m times and return the corresponding centers. The resulting empirical measure converges to ν as $m \to \infty$. In practice m is chosen to match the typical sample size of observations in the calibration set. It controls the numerical resolution of a candidate measure and is not a prediction of the future event count.

F Competing Baseline Regions

Note that predictions for each of our numerical results, tropical cyclones, and earthquake models have outputs as discrete probability grids that we represent by

$$P = (P_{ij}) \in [0,1]^{n_\phi \times n_\lambda}, \qquad \sum_{i=1}^{n_\phi} \sum_{j=1}^{n_\lambda} P_{ij} = 1,$$

defined on a latitude/longitude grid on S^2. Let

$$A = (A_{ij})$$

denote the cell area weights, with $A_{ij} > 0$. For an events at location (ϕ, λ), let $(i^\star, j^\star)$ be the nearest grid cell.

HDR Method The HDR (Hyndman, 1996) is obtained by sorting all cells by forecasts P_{ij} in decreasing order and only keeping the smallest collection of cells whose cumulative probability is at least $1 - \alpha$.

Let

$$\pi(1), \pi(2), \ldots, \pi(m)$$

be an ordering of the $m = n_\phi n_\lambda$ grid cells such that

$$P_{\pi(1)} \geq P_{\pi(2)} \geq \cdots \geq P_{\pi(m)}.$$

Define

$$K_{\text{HDR}}(\alpha) = \min \left\{ k : \sum_{\ell=1}^{k} P_{\pi(\ell)} \geq 1 - \alpha \right\}.$$

Then we have the HDR region

$$R_{\text{HDR},\alpha} = \{\pi(1), \ldots, \pi(K_{\text{HDR}}(\alpha))\}.$$

Minimum area (M1) The minimum area region ranks cells by probability mass per unit area. This method was influenced by Bayesian work (Box & Tiao, 1973)

$$D_{ij} = \frac{P_{ij}}{A_{ij}},$$

then as in HDR, keep the smallest number of cells with cumulative probability of at least $1 - \alpha$

Let

$$\pi_A(1), \ldots, \pi_A(m)$$

be an ordering such that

$$D_{\pi_A(1)} \geq D_{\pi_A(2)} \geq \cdots \geq D_{\pi_A(m)}.$$

Then define

$$K_{\text{MA}}(\alpha) = \min \left\{ k : \sum_{\ell=1}^{k} P_{\pi_A(\ell)} \geq 1 - \alpha \right\},$$

and the minimum area region is

$$R_{\text{MA},\alpha} = \{\pi_A(1), \ldots, \pi_A(K_{\text{MA}}(\alpha))\}.$$

This differs from HDR only by incorporating area into the equation. This method is identical if all cells have equal area.

Conformal HDR Min Area (M2) The conformal minimum area construction uses density score

$$D_{ij} = \frac{P_{ij}}{A_{ij}}.$$

For an observed event in cell (i, j), define the conformal score

$$S^{\mathrm{MA}}(i, j; P, A) = \sum_{r,s} P_{rs}\, \mathbf{1}\{D_{rs} > D_{ij}\}.$$

With calibration scores

$$S_t^{\mathrm{MA}} = S^{\mathrm{MA}}(i_t, j_t; P_t, A), \qquad t = 1, \ldots, n,$$

and conformal quantile $\hat{q}_\alpha$, the conformal minimum area region is

$$R_{\mathrm{cMA},\alpha} = \{(i, j) : S^{\mathrm{MA}}(i, j; P_{n+1}, A) \le \hat{q}_\alpha\}.$$

Note that conformal minimum area region is simply just the area weighted version of conformal HDR.

Diffusion (M3) We provide 3 generative baselines for point cloud generation. One drawback is that generative baselines we provide have no probabilistic guarantee that conformal inference and HDR methods give.

We implement a conditional denoising diffusion probabilistic model following the DDPM formulation Ho et al. (2020) directly on densities in $\mathbb{S}^2$. The forward process uses 100 diffusion steps with a linear variance from $\beta_1 = 0.001$ to $\beta_T = 0.02$. The denoiser is conditioned on the forecast through a permutation invariance point encoder and on the normalized diffusion time it uses hidden width 256 and predicts injected Gaussian noise. During training, the conditioning density is dropped with probability 0.2, and the objective combines the standard noise MSE with Chamfer reconstruction penalty on x_0 with weight $\lambda_{x_0} = 0.1$. We train for 300 epochs with Adam and a learning rate of 0.001, and batch size 8. Reverse sampling begins from Gaussian noise projected back to $\mathbb{S}^2$, uses the DDPM reverse updates with guidance scale 1. Thus the forward state is normalized to the sphere prior every reverse step. Completed densities are then projected back onto the sphere after the reverse chain is completed.

Spherical Flow Matching (M4) We implement a forecast conditioned spherical flow matching baseline for point clouds on $\mathbb{S}^2$. Each training example consists of a forecast cloud and its corresponding observed distribution. The forecast distribution is encoded using a permutation invariant point set encoder and used to condition the velocity field. For each minibatch, a source density is sampled uniformly from $\mathbb{S}^2$, and intermediate states are constructed by spherical linear interpolation (SLERP) between the source cloud and the observed target cloud. The velocity network combines the current point location, normalized time, and forecast conditioning representation, and uses two hidden layers of width 128 with $\tanh$ nonlinearities. Its output is projected onto the tangent space of $\mathbb{S}^2$ at the current point. Target velocities are approximated using a finite difference derivative of the SLERP path with step size 10^{-3} and are projected onto the tangent space. The model is trained by minimizing MSE between the predicted and target tangent velocities. We use Adam with learning rate 10^{-3}, minibatch size 32, and 200 training epochs. At generation time, the learned conditional velocity field is integrated using 100 forward Euler steps, with points renormalized to $\mathbb{S}^2$ after each step.

Variational Autoencoder (VAE) (M5) We additionally use a forecast conditioned variational autoencoder for spherical point clouds. Separate permutation invariant encoders map the forecast and target clouds to feature representations, which are combined to parameterize a 32 dimensional Gaussian latent variable. Both encoders and the residual decoder use hidden width 128. The decoder predicts a residual deformation of the conditioning forecast, scaled by 0.25, and the resulting points are normalized to $\mathbb{S}^2$. The training objective consists of a Chamfer reconstruction loss, a KL divergence penalty with weight 0.001, and a repulsion penalty with weight 0.05 to discourage point collapse. We train using Adam with learning rate 10^{-3}, batch size 16, and 200 epochs. At generation time, latent variables are sampled from a standard Gaussian and decoded conditionally on the forecast cloud.

G Simulated Densities

G.1 Additional Scoring

We provide all results, real data and simulations, in terms of Euclidean energy score.

Each empirical density is first converted into a probability mass vector on a shared spatial grid and then the standard ensemble energy score is computed using Euclidean distance within that vector space. Suppose Y is our observed empirical density, and $Z_1, \cdots, Z_M$ are the forecast ensemble densities, then we define Euclidean energy score as

$$ES_{Euc} = \frac{1}{M}\sum_{m=1}^{M} ||p(Z_m) - p(Y)||_2 - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{l=1}^{M} ||p(Z_m) - p(Z_l)||_2$$

where $p(Y), p(Z_m) \in \mathbb{R}^K$ where K is the corresponding number of grid cells. We use a 72×36 grid for tropical cyclone genesis locations, and a $0.25°$ California grid for the earthquake data evaluation.

Table 4: Euclidean energy score on consistently gridded probability mass vectors. Values are reported as mean (standard error) over the held out test realizations. Lower is better. Euclidean energy score is multiplied by 10.

Process	M1	M2	M3	M4	M5	VF	CVF
UWN	0.5198 (0.0509)	0.5184 (0.0511)	0.5056 (0.0524)	0.5047 (0.0523)	0.5057 (0.0524)	0.4470 (0.0395)	0.4496 (0.0414)
Spiral	0.1099 (0.0005)	0.1090 (0.0004)	0.1038 (0.0004)	0.1042 (0.0004)	0.1048 (0.0004)	0.1017 (0.0004)	0.1018 (0.0004)
3-Mode	0.1031 (0.0003)	0.1027 (0.0003)	0.1016 (0.0003)	0.1028 (0.0003)	0.1015 (0.0003)	0.1019 (0.0003)	0.1020 (0.0003)
Hawkes	0.0765 (0.0016)	0.0762 (0.0015)	0.0753 (0.0016)	0.0754 (0.0015)	0.0752 (0.0016)	0.0752 (0.0015)	0.0753 (0.0015)

Table 5: Euclidean energy score (standard error) for both data applications multiplied by 100.

Dataset	M1	M2	M3	M4	M5	VF	CVF
Tropical Cyclones	9.24 (0.12)	9.36 (0.14)	9.12 (0.16)	9.84 (0.14)	9.08 (0.14)	**8.46 (0.12)**	8.58 (0.11)
Earthquakes	27.74 (2.07)	27.39 (2.10)	33.60 (1.83)	26.35 (2.14)	32.67 (1.85)	24.61 (1.80)	**23.95 (1.83)**

G.2 White Noise on a Sphere

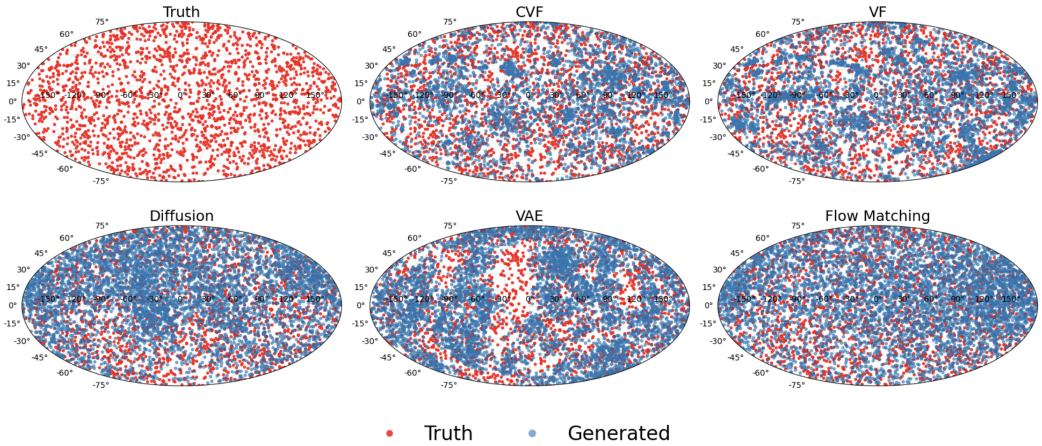

Figure 5: Homogeneous Poisson model with generated sample prediction

We simulate white noise on a sphere with a homogeneous Poisson process

$$N \sim Poisson(4\pi\lambda)$$

where $\lambda > 0$ is a constant intensity per unit area and N is the total count. We condition on N and points are i.i.d. uniform on the sphere. In our experiment, we use $\lambda = 20$

G.3 MISSPECIFIED SPIRAL

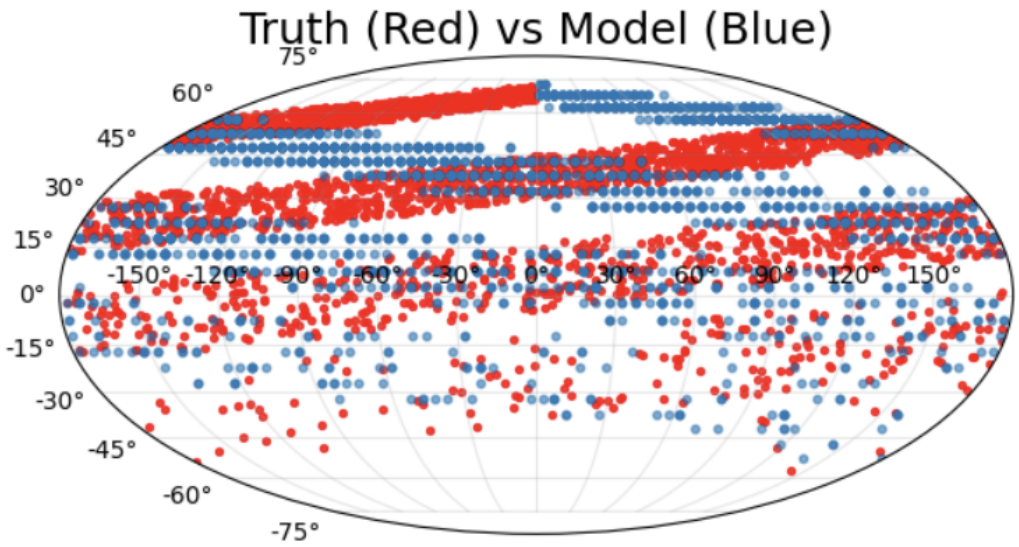

Figure 6: Spiral simulation with opposite direction forecast model. Both have exactly the same features, but inverted to learn a very difficult transition along with a specified area points can and cannot lie.

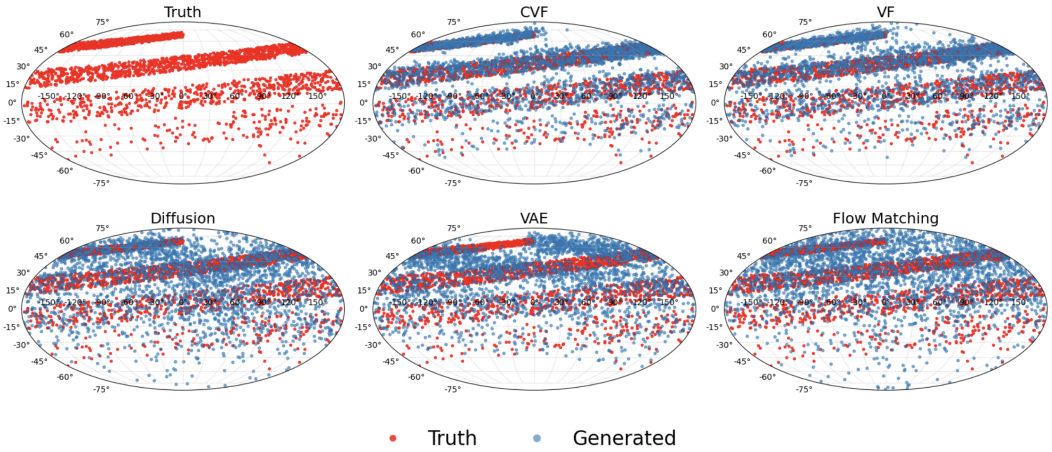

Figure 7: Spiral simulation with opposite direction oracle model fit (Top Left)

The spiral simulation is built on the sphere using a spiral curve structure. This simulation also has a misspecified model where the spiral spins the opposite direction and has higher rate near the top of the sphere. To generate points along the spiral, the parameter t is sampled from a beta distribution with parameters $\alpha = 3$ and $\beta = 1.2$. This process places points on expanding bands as they travel to the bottom of the sphere. Each replicate in the spiral contains a true cloud of 50 points and has 50 timepoints to mimic smaller data with more difficult manifold structure.

G.3.1 POINTS INSIDE BAND

Due to the spiral simulation having exact band constraints, we can check how many individual points lie within these tight bands. Notice that CVF keeps a majority of test points within the band, while other methods have a harder time transporting points with conditional residuals, or injected information from the misspecified model.

Table 6: Prediction set points within misspecified spiral test set band.

Method	% Points in Test Band
Diffusion	53.36
Flow Matching	44.72
VAE	46.04
VF	83.12
CVF	83.48

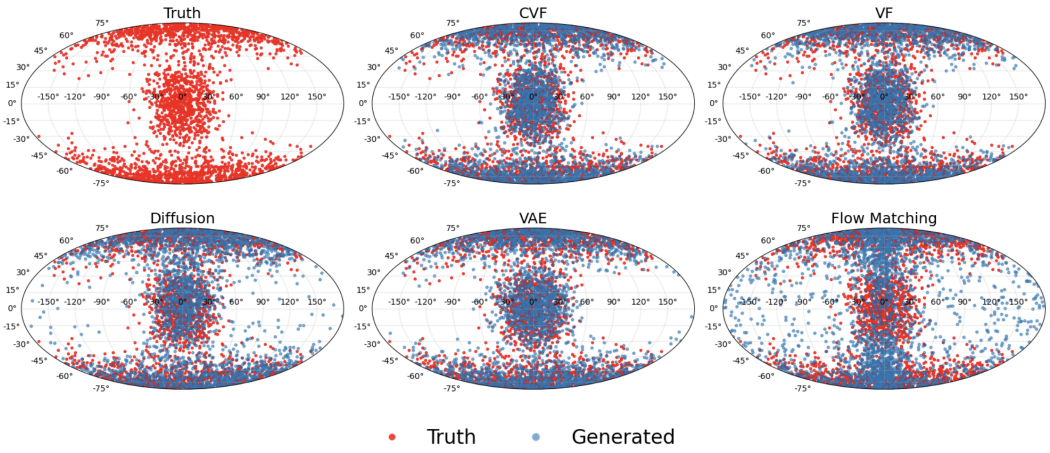

Figure 8: 3 Mode distribution with oracle model fit (Top Left)

G.4 Three Mode Density

Another important aspect that our method must show is high performance for point processes that have structured, non-uniform behavior unlike homogeneous Poisson processes (Baddeley et al., 2015). We consider one 3 mode density with a specified intensity function to represent three different modes on the globe. The core of this density utilizes aspects of the von Mises Fisher distribution in parallel with that of an inhomogeneous Poisson process which is widely used in spatial statistics on the globe (Banerjee et al., 2005).

We define 3-Mode as a sample from an inhomogeneous Poisson process on a "belt" defined as $B(x) = e^{-(\frac{z}{\sigma})^2}$ placing the centers at the central meridian and three modes defined as $H(x) = \sum_m^M e^{\kappa_h \mu_m' x}$ This leads to the defined intensity function

$$\lambda(x) = \lambda_0(1 + be^{\frac{-z^2}{\sigma_b^2}} + a_h \sum_m^M e^{\kappa_h \mu_m' x})$$

To sample we use thinning from a uniform distribution with the conservative upper bound of λ_{max} We utilize parameters

$$\lambda_0 = 0.01,\ b = 0.6\ \sigma_b = 0.3,\ M = 3,\ a_h = 1,\ \kappa_h = 10$$

G.5 Hawkes Process

Hawkes processes are "self exciting" point processes that were designed to have inhomogeneous Poisson process randomness, but to have generated points have larger probabilities to generate more points (Hawkes, 1971). Specifically in Earthquake modeling, the idea of utilizing Hawkes based processes for simulation or modeling is popular (Ogata, 1988). Due to our interest in recreation of structures occurring due to a structure on a sphere, we create an equatorial belt utilizing a Hawkes process where events near the equator are more common and slowly taper out as the distance increases with a self exciting property.

We define Hawkes process $Hawkes$ on $[0, T] \times S^2$ with temporal Hawkes intensity

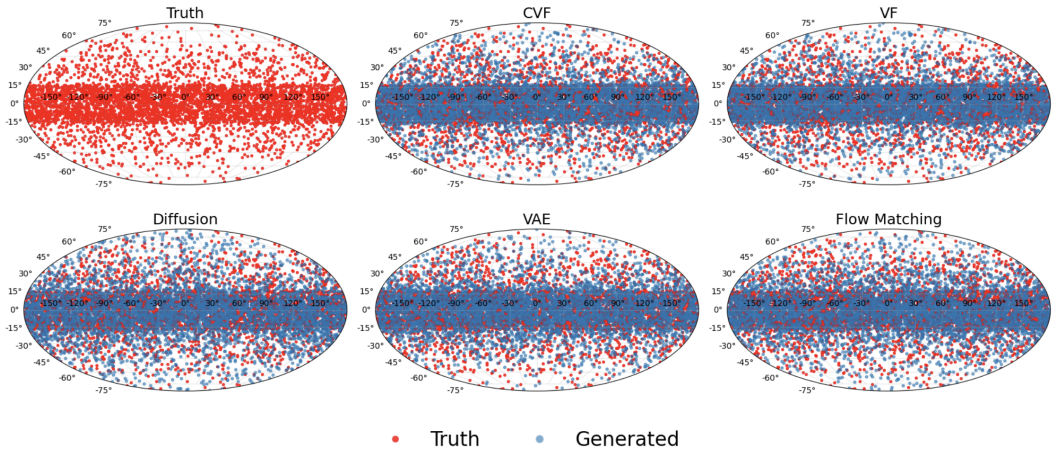

Figure 9: Hawkes process on a belt with oracle model fit (Top Left)

$$\lambda(t|H_t) = \mu + \sum_{T_i<t} \alpha e^{-\beta(t-T_i)};\ t \in [0,T]$$

We then include type probabilities

$$P(immigrant|H_t) = \frac{\mu}{\lambda(t)},\ P(Offspring|H_t) = \frac{c_j(t)}{\lambda(t)}$$

In addition, with the same intensity, we define a "belt" region as $B = \{x \in S^2 : |x_3| < z_{max}\}$ where $x = \{x_1, x_2, x_3\}$ and $z_{max} \in (0,1)$. We let g_{belt} be uniform on the belt and the immigrant points have locations

$$X|immigrant \sim uniform(belt)$$

and offspring as

$$X|offspring \sim vMF(\mu = X_j, \kappa_{space})$$

We utilize parameters

$$\mu = 0.7, \alpha = 0.6, \beta = 1.5, \kappa_{space} = 12, z_{max} = 0.3$$

G.6 Tropical Cyclone Model

Table 7: Tropical cyclone energy distance and manifold distance metrics. Standard errors are reported in parentheses.

Metric	M1	M2	M3	M4	M5	VF	CVF
ED	3.892 (2.185)	4.096 (2.292)	0.881 (0.527)	1.617 (0.932)	0.628 (0.971)	0.419 (0.434)	0.530 (0.463)
Pred–Pred	4.910 (-)	4.035 (-)	5.602 (-)	5.545 (-)	6.815 (-)	4.823 (-)	4.730 (-)
True–Pred	6.942 (1.066)	6.606 (1.122)	5.783 (0.511)	6.122 (0.618)	6.263 (0.570)	5.162 (0.493)	5.171 (0.506)
MD	1.449 (0.0333)	1.491 (0.0327)	1.787 (0.0543)	2.341 (0.0589)	1.807 (0.0500)	1.651 (0.0881)	0.928 (0.459)

We construct a frozen presence background model for tropical cyclone genesis prediction. We utilize environmental data obtained from the Copernicus data store along with cyclone genesis data from the IBTrACS database for the first time a storm attains tropical cyclone status. For each year for the training period (1980-1989), the genesis locations are mapped onto the same forecast grid as a distribution. This is then smoothed and normalized to form a yearly target probability map for downstream analysis.

The purpose of this model is to produce a probability distribution over the globe to estimate the spatial distribution of genesis locations. This fixed design will allow us to not violate exchangeability for conformal methods.

Let

$$X_y \in \mathbb{R}^{C\times H\times W}$$

denote the yearly environmental tensor for year y by stacking 12 predictor fields across all variables. Let

$$\mathcal{G}_y = \{(i_k, j_k)\}_{k=1}^{N_y}$$

denote the set of observed genesis cells in year y. For each forecast year y prior to 1990, we form a training dataset from all years $y' < y$. Any positive values are taken from cells in $\mathcal{G}_y$ while the negative are drawn from randomly sampled background cells. The vector associated with a grid cell is the channel vector

$$X_{y'}[:, i, j]$$

We fit a regularized logistic regression model on this dataset interpreting the fitted output as a relative occurrence intensity.

We let $z_r = 1$ if row r contained a genesis location or cell, and $z_r = 0$ if row r corresponds with a background cell. We then fit l_2 regularized logistic regression with conditional probability

$$P(z_r = 1|\tilde{x}_r) = \frac{1}{1 + exp(-(b + \tilde{x}_r'\beta))}$$

The fitted parameters are simply obtained by minimizing the regularized logistic loss. This model is held out for all later forecast years. For each year, we evaluate the fitted model at every cell on the globe to obtain

$$\hat{p}_y(i, j) = \frac{1}{1 + exp(-(b + \tilde{x}'_{y,i,j}\beta))}$$

and normalize to

$$\hat{\pi}_y(i, j) = \frac{\hat{p}_y(i, j)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\hat{p}_y(i, j)}$$

as the forecast probability map for year y.

For conformal methods, we hold out the years 1990-2010 as calibration with no additional training to the model. SSW scores are recorded for VF and CVF methods. Our data typically has between 50 and 110 points a year.

Table 8: Temporal diagnostics for the tropical cyclone conformity scores.

Diagnostic	Statistic	p-value
Spearman time trend	$\rho = 0.032$	0.856
Ljung–Box (lag 5)	$Q(5) = 7.22$	0.205
Cramér–von Mises	$T = 0.026$	0.998
Lag-1 correlation	$r_1 = -0.166$	0.357

G.7 EARTHQUAKES

Table 9: California earthquake energy distance and manifold distance metrics. Standard errors are reported in parentheses.

Metric	M1	M2	M3	M4	M5	VF	CVF
ED $\times 100$	0.315 (0.541)	0.322 (0.582)	0.762 (0.443)	0.501 (0.654)	0.543 (0.479)	0.116 (0.334)	0.145 (0.369)
Pred–Pred	1.041 (-)	1.067 (-)	1.294 (-)	1.069 (-)	1.102 (-)	1.829 (-)	1.825 (-)
True–Pred	1.522 (0.311)	1.539 (0.332)	1.872 (0.257)	1.629 (0.354)	1.667 (0.286)	1.817 (0.192)	1.829 (0.202)
MD $\times 100$	0.531 (0.401)	0.677 (0.325)	1.382 (0.254)	0.987 (0.369)	1.121 (0.228)	0.779 (0.361)	0.536 (0.473)

We model earthquake occurrences in the region of California using a grid spatial forecast model constructed from historical data. Let the past earthquake data be

$$\mathcal{D} = \{(t_i, l_i^{lat}, l_i^{lon}, m_i, d_i)\}_{i=1}^{N}$$

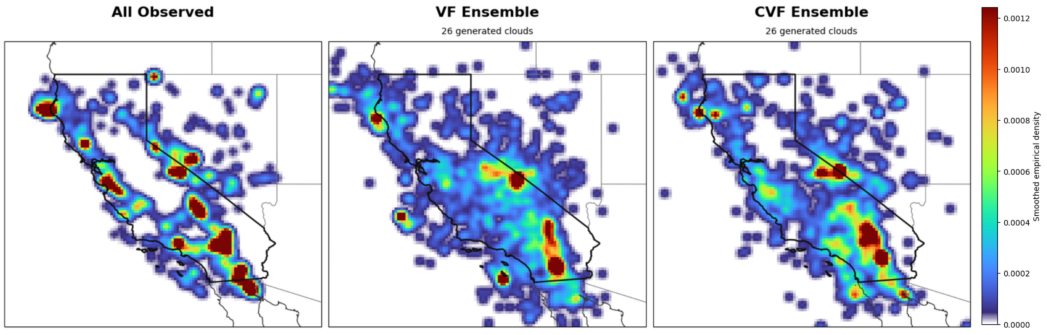

Figure 10: Earthquake velocity field and constrained velocity field with all observed Earthquakes to show entire support.

where t_i is the time of the event, (l_i^{lat}, l_i^{lon}) is the earthquake location, m_i is the magnitude, and d_i is the depth of the earthquake. We restrict data to $m_i > 4.0$ in the California region, denoted by

$$\mathcal{R} = [l_{min}^{lon}, l_m^{lon}ax] \times [l_{min}^{lat}, l_{max}^{lat}]$$

We discretize this on a fixed latitude longitude grid and let each grid contain cells $\mathcal{C}_{jk}$ indexed by the latitude bin j and longitude bin k with center c_{jk}.

Instead of re-fitting the model each year, we instead construct a fixed spatial forecast from an initial training block in order to support the exchangeability assumption. We fit the model once using earthquakes from 1976 to 1984, use the years 1985 through 2010 for calibration, and evaluate on the 2011 through 2025 data. The observed earthquakes in each calibration or test year are then compared against samples from the fixed forecast grid.

The fitted forecasting model is simply a cumulative two dimensional histogram predictor. For each gridcell $\mathcal{C}_{jk}$, we count the number of historical earthquakes that fall into that cell

$$N_{jk}(T) = \sum_{i=1}^{N} 1\{t_i < T, m_i \geq 4.0, (l_i^{lat}, l_i^{lon}) \in \mathcal{C}_{jk}\}$$

$N_{jk}(T)$ defines an unnormalized spatial intensity which is then converted to a probability surface by normalization. This gives a discrete forecast distribution over the grid of California representing an empirical gridded version of an adaptive smoothed seismicity model typically used in forecasts. Our data typically has 11 to 140 points per yearly point cloud.

Table 10: Diagnostics for the 2011–2025 held-out earthquake conformity scores.

Diagnostic	Statistic	p-value
Spearman trend	$\rho = 0.275$	0.082
Ljung–Box, lag 5	$Q(5) = 8.177$	0.147
Cramér–von Mises	$T = 0.112$	0.551
Lag-1 correlation	$r_1 = 0.005$	0.974

G.8 Variance and Sample Diversity

We now evaluate the effectiveness of sampling via the joint velocity field (Equation 16) for representing the manifold constrained prediction sets in practice (9). Figure 11 shows the distribution of SSW distances between actual density at the 90th quantile and between our models samples at 90th quantile and the actual data. In each case the model is able to simulate empirical densities that have similar variability with the data as the data does with itself.

G.9 Chicago Crime Data

As an additional dataset benchmark, we consider reported crime incidents in Chicago using the *Crimes–2001 to Present* dataset from the City of Chicago Data Portal (Chicago Police Department,

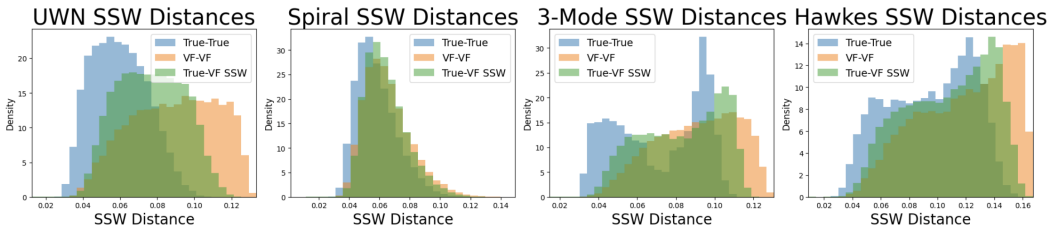

Figure 11: Distribution of intrasample distances from CVF method ($\alpha = 0.1$) compared with intersample distances of the data and the CVF method (90th percentile)

2026). The data are provided by the Chicago Police Department and contain reported crime incidents extracted from the department's CLEAR system, together with incident dates and geographic coordinates. We use incidents occurring from January 2022 through December 2025 with nonmissing latitude and longitude and restrict the analysis to the geographic region $41.60° \leq \text{latitude} \leq 42.05°$ and $-88.00° \leq \text{longitude} \leq -87.45°$. This data is much richer than tropical cyclone genesis and earthquake locations, providing a different type of data for analysis.

Incidents are aggregated into Monday start weekly intervals. For each week, we construct an empirical spatial probability distribution on a $0.01° \times 0.01°$ latitude–longitude grid by counting incidents within grid cells and normalizing the resulting histogram to unit probability mass. Let p_t denote the normalized spatial distribution for week t. For each week, we construct a simple rolling forecast

$$\widehat{p}_t = \frac{1}{52} \sum_{j=1}^{52} p_{t-j}.$$

Importantly, each historical weekly histogram is normalized before averaging, so weeks with larger numbers of reported incidents do not receive greater weight in the forecast. This construction uses only observations preceding the forecast week and therefore provides a simple nonparametric spatial forecasting benchmark. After the initial lookback period, the procedure produces 158 weekly forecast to observation pairs. For the distributional comparisons, both forecasts and observations are represented by point clouds of cardinality 500, and latitude/longitude coordinates are mapped to their corresponding unit vectors on $\mathbb{S}^2$.

To reduce sensitivity to an arbitrary chronological split while preserving local temporal structure, calibration and evaluation weeks are assigned using contiguous four week blocks. Blocks are allocated separately within each calendar year, with approximately two thirds assigned to calibration and one third to evaluation, so that both sets contain observations from comparable temporal regimes.

We assess empirical conformal coverage over 50 repeated randomized year balanced blocked splits.

Table 11: Empirical SSW coverage over 50 randomized year balanced temporally blocked splits of the Chicago crime data.

Nominal Coverage	Mean Empirical Coverage	SD
90%	91.04%	4.89%
85%	85.82%	6.76%
80%	81.05%	8.11%
75%	77.02%	9.24%
70%	72.37%	9.18%

Coverage is close to the nominal level across the repeated blocked allocations. Since the weekly observations may exhibit temporal dependence, these results are interpreted as an empirical robustness assessment rather than as an application of the standard i.i.d. finite sample conformal guarantee.

G.10 Chicago Crime Results

For the comparison of M1 through M5, VF, and CVF, we use a single prespecified blocked split. The generative baselines M3 through M5 are trained on paired forecast and observed point clouds from

Table 12: Chicago crime SSW energy distance and manifold distance metrics, both scaled by 10^5, with standard errors in parentheses.

Metric	M1	M2	M3	M4	M5	VF	CVF
ED$\times 10^5$	1.44 (0.180)	1.49 (0.208)	130.78 (0.291)	50.59 (0.499)	42.96 (0.236)	0.21 (0.097)	1.15 (0.125)
MD$\times 10^5$	0.28 (0.007)	0.28 (0.010)	49.83 (0.149)	3.53 (0.029)	19.21 (0.194)	0.38 (0.031)	0.42 (0.009)

the calibration portion and are then conditioned on the held out weekly forecasts. In addition, for generation of prediction ensembles, the white noise generation is restricted to the Chicago region: $41.60^\circ \leq \text{latitude} \leq 42.05^\circ$ and $-88.00^\circ \leq \text{longitude} \leq -87.45^\circ$. M1 and M2 are constructed from the corresponding forecast probability grids. VF and CVF use the same calibration forecast to observation pairs and SSW calibration scores, ensuring that all methods are evaluated on the same held out weeks. We evaluate the resulting prediction regions using SSW energy distance and the manifold discrepancy described in Section 3.2.

The Chicago crime application shows that VF based methods can be applied, without change, to a specific small region on the sphere. The choice of manifold distance in this paper was chosen for densities with less calibration and fewer points per density. Due to the dense nature of the Chicago dataset, the manifold penalty was not needed and was not as effective as a simple non penalized velocity flow. However, testing of different manifold penalties could produce desirable results.