SPHERICAL HARMONIC OPTIMAL TRANSPORT: APPLICATION TO CLIMATE MODELS COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal transport provides a powerful framework for comparing measures while respecting the geometry of their support, but comes with an expensive computational cost, hindering its potential application to real world use cases. On manifolds, convolutional algorithms based on the heat kernel have been proposed to alleviate this cost, but their theoretical properties remain largely unexplored. We establish that the heat kernel cost converges to the optimal transport cost as time vanishes in the balanced and unbalanced cases. In the specific case of the 2-sphere $\mathbb{S}^2$, we ensure that the associated Sinkhorn divergences retains the desirable geometric and analytic properties of classical optimal transport discrepancies. Moreover, we leverage the harmonic structure of the sphere to derive a fast Sinkhorn algorithm, requiring only $\mathcal{O}(n)$ memory and $\mathcal{O}(n^{3/2})$ time per iteration, with fully dense GPU-friendly operations. We validate its computational efficiency on synthetic data, and discuss its potential use in the evaluation of global climate models, providing both spatial and seasonal insights into models performances.

1 INTRODUCTION

Numerical OT. Optimal transport (OT) provides a principled framework for comparing probability measures that explicitly accounts for the geometry of their support, and has become a central tool in machine learning (Peyré et al., 2019), with applications spanning generative modeling (Arjovsky et al., 2017), domain adaptation (Courty et al., 2017), and barycenter estimation (Cuturi, 2013; Peyré et al., 2019).

While data is often assumed to live in Euclidean spaces, many situations exhibit phenomena that inherently lie on curved manifolds. A canonical example is the two-dimensional sphere $\mathbb{S}^2$, which arises naturally in directional statistics (Mardia & Jupp, 2000) and geospatial or climate sciences (Garrett et al., 2024; Hyun et al., 2022). This has motivated the development of OT methods that operate directly on the sphere and respect its intrinsic Riemannian geometry (Bonet et al., 2023; Liu et al., 2024).

A major algorithmic advance was the introduction of entropic regularization, leading to the Sinkhorn algorithm and enabling scalable OT approximations (Cuturi, 2013). Entropic regularization also gives rise to the Sinkhorn divergence, which corrects biases inherent to the regularized problem and yields a principled loss function for machine learning under suitable assumptions on the kernel (Feydy et al., 2019). In Euclidean spaces with quadratic cost, the underlying Gibbs kernel is Gaussian and enjoys well-understood analytical properties. On curved manifolds such as $\mathbb{S}^2$, however, the Gibbs kernel generally fails to be positive definite (Jayasumana et al., 2015; Feragen et al., 2015; Da Costa et al., 2025), invalidating these desirable properties. Beyond these theoretical limitations, the most computationally efficient Sinkhorn implementations, such as the recent FLASHSINKHORN (Ye et al., 2026), exploit the inner-product structure of the squared Euclidean cost, a decomposition unavailable for the squared geodesic distance $d(x, y) = \arccos(\langle x, y\rangle)^2$ on $\mathbb{S}^2$. On a different level, sliced Wasserstein methods are traditional scalable alternatives. They have been extended to the sphere by projecting onto great circles (Bonet et al., 2023) and further explored with alternative slicing techniques (Quellmalz et al., 2023; 2024; Tran et al., 2024) or linearised variants (Liu et al., 2024). While efficient at handling large number of samples, they usually do not provide transport plans, are subject to randomness via their Monte-Carlo approximations and their unbalanced versions Bonet et al. (2024) were never considered on the sphere.

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.

An appealing alternative, proposed in the context of convolutional OT (Solomon et al., 2015), is to replace the Gibbs kernel with the heat kernel of the underlying manifold, motivated by Varadhan's formula linking heat diffusion to geodesic distances (Varadhan, 1967). Yet this approach remains theoretically and computationally underdeveloped on $\mathbb{S}^2$: the conditions guaranteeing well-posedness of the resulting Sinkhorn divergence have not been verified for the heat kernel on the sphere, and existing implementations rely on generic numerical solvers that discard its rich harmonic structure.

OT for Climate models Comparison. One particular and exciting applications of OT on the two-dimensional sphere is the evaluation of global climate models (GCMs). Evaluating the realism of GCMs is a central challenge in climate science. The sixth-generation Coupled Model Intercomparison Project (CMIP6) (Eyring et al., 2016) provides a standardized ensemble of model simulations that can be compared against observational references. A key difficulty is that geophysical fields such as precipitation are measures that are spatially structured, non-negative, and naturally defined on the globe $\mathbb{S}^2$, yet standard evaluation metrics such as Root Mean Square Error (RMSE) or pattern correlation treat them as flat Euclidean fields, ignoring the spherical geometry and the distributional nature of precipitation. Recently, OT-based metrics have been proposed to address these limitations and provide a more holistic assessment of model performance (Vissio et al., 2020; Garrett et al., 2024; Veiga Rodrigues & Odderskov, 2025; Xie et al., 2025). However, these works have relied on balanced OT formulations that do not account for the systematic global mean biases exhibited by GCMs relative to observations, and have not fully leveraged the geometry of the sphere as they mostly work on time series 1D signals (Garrett et al., 2024; Xie et al., 2025). Our work addresses these gaps by developing an unbalanced OT (UOT) framework based on the heat kernel for evaluating GCMs on $\mathbb{S}^2$, providing both theoretical guarantees and practical tools for this important application.

Contributions. Our contributions can be summarized as follows:

- We study entropically regularized balanced and unbalanced OT and Sinkhorn divergences. Using the heat kernel as a replacement for the Gibbs kernel associated with the squared geodesic distance, we show that this cost converges to the optimal transport cost in the small-time limit. In the case of the 2-sphere, we moreover show that it satisfies the key properties required by the Sinkhorn divergence framework.
- We exploit the spherical harmonic transform to derive SHOT, an efficient Sinkhorn algorithm for measures on $\mathbb{S}^2$ via fast convolutions against the heat kernel, enabling scalable computation of Sinkhorn divergences, with performance validated on synthetic data.
- We apply our framework to the evaluation of global climate models, comparing the precipitation fields of CMIP6 models against ERA5 reanalysis data. We show that our unbalanced OT approach captures both global and regional biases in model performance, and provides insights into seasonal differences in model skill across different latitudinal bands.

2 Background

Let $\mathcal{M}$ be a compact, connected Riemannian manifold and $C : \mathcal{M}^2 \to \mathbb{R}_+$ be a measurable cost function, which we will typically take to be the squared geodesic distance d^2 on $\mathcal{M}$.

Definition 2.1 (Optimal Transport)**.** Let P and Q be two probability measures on $\mathcal{M}$. The *optimal transport* (OT) problem is

$$\mathrm{OT}_C(P, Q) := \inf_{\pi \in \mathcal{U}(P,Q)} \int_{\mathcal{M}^2} C(x, y)\, \mathrm{d}\pi(x, y). \tag{1}$$

where $\mathcal{U}(P, Q)$ is the set of *couplings* between P and Q, i.e., joint probability measures on $\mathcal{M}^2$ with marginals P and Q.

Existence of optimal couplings is well understood (Villani, 2009; Santambrogio, 2015). The choice $C = d^2$ gives $\mathrm{OT}_{d^2} = W_2^2$, the squared 2-Wasserstein distance. While the OT problem (1) is convex, it is typically not strictly convex and may admit multiple optimal couplings, which complicates both analysis and computation. Moreover, solving (1) exactly between two discrete measures with support of size n requires $\mathcal{O}(n^3)$ operations (Peyré et al., 2019), which is prohibitive at scale. A further limitation is that (1) requires P and Q to have equal total mass, a constraint often violated in practice

when comparing unnormalized densities, point clouds of different sizes, or measures corrupted by outliers. Both issues are fixed by the following formulation.

Definition 2.2 (Entropic Unbalanced OT, Chizat et al., 2018)**.** For $\varepsilon \geq 0$ and $\tau \in (0, +\infty]$, the *entropic unbalanced OT* between two positive Radon measures P and Q on $\mathcal{M}$ is

$$\mathrm{UOT}^{\tau}_{C,\varepsilon}(P,Q) := \inf_{\pi \in \mathcal{R}_+(\mathcal{M}^2)} \int_{\mathcal{M}^2} C \,\mathrm{d}\pi + \varepsilon \,\mathrm{KL}(\pi \mid P \otimes Q) + \tau \left(\mathrm{KL}(\pi^1 \mid P) + \mathrm{KL}(\pi^2 \mid Q)\right), \quad (2)$$

where $\mathcal{R}_+(\mathcal{M}^2)$ is the set of positive Radon measures on $\mathcal{M}^2$, KL is the Kullback-Leibler divergence (with the convention $+\infty \cdot \mathrm{KL}(\cdot \mid \cdot) = \iota_{\{0\}}(\mathrm{KL}(\cdot \mid \cdot))$) and π^1 and π^2 are the first and second marginals of π, respectively.

The parameter τ interpolates between two regimes. At $\tau = +\infty$, exact marginal constraints are enforced, yielding *balanced* entropic OT (and (1) as $\varepsilon \to 0$). For finite $\tau > 0$, the problem becomes *unbalanced*: mass creation or destruction is allowed at a cost proportional to KL divergence from each marginal, with $\tau \to 0$ recovering unconstrained transport. We write $\mathrm{OT}_{C,\varepsilon} := \mathrm{UOT}^{+\infty}_{C,\varepsilon}$ for the balanced case and $\mathrm{UOT}^{\tau}_{C} := \mathrm{UOT}^{\tau}_{C,0}$ for the unregularized unbalanced OT. A key advantage of (2) is that it admits an efficient Sinkhorn iterative solver with cost $\mathcal{O}(n^2)$ for n support points.

Sinkhorn Divergence. Although computationally efficient, entropic regularization introduces a systematic bias: in general, $\mathrm{UOT}^{\tau}_{C,\varepsilon}(Q,Q) \neq 0$. The Sinkhorn divergence (3) corrects this bias by subtracting the self-transport terms.

Definition 2.3 (Séjourné et al., 2019)**.** For $\varepsilon > 0$, the *Sinkhorn divergence* (SD) between $P, Q \in \mathcal{R}_+(\mathcal{M})$ is defined as

$$S^{\tau}_{C,\varepsilon}(P,Q) \coloneqq \mathrm{UOT}^{\tau}_{C,\varepsilon}(P,Q) - \tfrac{1}{2}\left(\mathrm{UOT}^{\tau}_{C,\varepsilon}(P,P) + \mathrm{UOT}^{\tau}_{C,\varepsilon}(Q,Q)\right) + \tfrac{\varepsilon}{2}(m(P) - m(Q))^2, \quad (3)$$

where $m(\cdot)$ denotes the total mass of the measure on the manifold $\mathcal{M}$.

The SD inherits desirable theoretical properties: it defines a symmetric, positive definite, smooth loss function that is convex in each of its input variables and metrizes the convergence in law when some conditions on the space and the cost function are satisfied (Séjourné et al., 2019, Theorem 5 & 6).

Convolutional Optimal Transport Every iteration of the Sinkhorn algorithm requires the computation of the convolution of a function $f \in L^2(\mathcal{M})$ against $\mathcal{K}_{C,\varepsilon}(x,y) = \exp(-C(x,y)/\varepsilon)$. For a general kernel $K : \mathcal{M}^2 \to \mathbb{R}$, the convolution is defined as the integral operator

$$Kf(x) = \int_{\mathcal{M}} K(x,y)\, f(y) \,\mathrm{d}\sigma(y). \quad (4)$$

Although the straightforward way to compute it has a quadratic complexity in the size of the discretization, more efficient algorithms exist in some situations. Indeed, in Euclidean space, the kernel $\mathcal{K}_{d^2,\varepsilon}$ is the standard Gaussian kernel, in such a way that each iteration only requires a simple Gaussian convolution, which can be computed with complexity $\mathcal{O}(n \log n)$ for a grid of size n.

On a general Riemannian manifold, the Gibbs kernel $\mathcal{K}_{d^2,\varepsilon}$ has a very similar role, smoothing functions according to the geodesic distance. However a fast implementation of the convolution against it does not exist. Solomon et al. (2015) sidestep this difficulty by approximating $\mathcal{K}_{d^2,\varepsilon}$ with the heat kernel.

For $t > 0$, the *heat kernel* $\mathcal{H}_t : \mathcal{M}^2 \to \mathbb{R}$ is defined by the property that, for every $y \in \mathcal{M}$, the function $(t,x) \mapsto \mathcal{H}_t(x,y)$ is the minimal positive solution of the heat equation $\partial_t u(t,x) = \Delta u(t,x)$, with initial condition $u(0,x) = \delta_y(x)$ in the sense of distributions, and where Δ is the Laplace-Beltrami operator on $\mathcal{M}$. The heat kernel is smooth, symmetric, and, as $\mathcal{M}$ is complete, strictly positive for all $t > 0$ (Grigoryan, 2009, Thm. 7.20, Cor. 8.12).

The approximation of the Gibbs kernel with the heat kernel is justified by Varadhan's asymptotic formula (Varadhan, 1967), which implies that $\mathcal{H}_{\varepsilon/4} \approx \mathcal{K}_{d^2,\varepsilon}$ for small ε. Thus, we can approximate $\mathrm{OT}_{d^2,\varepsilon}$ by replacing the Gibbs kernel with $\mathcal{H}_{\varepsilon/4}$ in the convolutions involved in Sinkhorn iterations.

With this substitution, each Sinkhorn iteration amounts to diffusing a function over $\mathcal{M}$ for a short time interval $\varepsilon/4$. On meshes or two-dimensional manifolds, heat diffusion can be discretized

via an implicit Euler step, reducing each iteration to solving a sparse linear system involving the discrete Laplace-Beltrami operator. Pre-factoring this system via a sparse Cholesky decomposition amortizes the cost across iterations. However, sparse Cholesky is not GPU friendly, as its computation follows an irregular dependency structure which limits parallelism. Also, its sparse memory access patterns lead to poor memory coalescing, preventing efficient use of GPU bandwidth and throughput, hindering its use in large scale OT problems.

3 Spherical Harmonic Optimal Transport

3.1 Theoretical framework

Heat kernel and UOT. We now formalize the connection between heat kernel-regularized OT and classical OT. For $\varepsilon > 0$, define the smooth *time-ε-cost* $\widehat{C}_\varepsilon := -\varepsilon \log \mathcal{H}_{\frac{\varepsilon}{4}}$. The heuristic approximation proposed by Solomon et al. (2015) suggests that $\mathrm{OT}_{d^2,\varepsilon}$ can be approximated by $\mathrm{OT}_{\widehat{C}_\varepsilon,\varepsilon}$. Our main result establishes that this approximation becomes exact as $\varepsilon \to 0$ in the balanced and unbalanced cases. This is not a direct consequence of standard entropic OT convergence results (e.g. those of (Carlier et al., 2017; Nutz, 2021; Carlier et al., 2023)), as the cost function itself varies with ε.

Theorem 3.1. *For all $P, Q \in \mathcal{R}_+(\mathcal{M})$ and $\tau \in (0, +\infty]$ we have*

$$\lim_{\varepsilon \to 0} \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}(P, Q) = \mathrm{UOT}^\tau_{d^2}(P, Q).$$

The proof is provided in the appendix A.1. This rigorously justifies heat kernel-based Sinkhorn as an efficient proxy for classical OT. We note $\hat{S}^\tau_\varepsilon := S^\tau_{\widehat{C}_\varepsilon,\varepsilon}$ for brevity ($\hat{S}_\varepsilon$ in the balanced case).

The following result justifies the use of the Sinkhorn divergence theory in our setting *i.e.* $\mathcal{M} = \mathbb{S}^2$, ensuring $\hat{S}^\tau_\varepsilon$ retains its desirable properties. The proof is deferred to Appendix A.2.

Proposition 3.2. *For any $\varepsilon > 0$ and $\tau \in (0, \infty]$, $\hat{S}^\tau_\varepsilon$ defines a symmetric, positive definite, smooth function that is convex in each of its input variables and metrizes the convergence in law.*

Taken together, Theorem 3.1 and Proposition 3.2 provide a solid theoretical foundation for heat kernel-regularized OT on $\mathbb{S}^2$. We therefore propose to leverage spherical harmonics, which diagonalize the Laplace-Beltrami operator on $\mathbb{S}^2$ and reduce heat diffusion to pointwise multiplication in frequency space, enabling fully parallelizable, GPU-friendly Sinkhorn iterations.

Spherical Harmonics. Functions in $L^2(\mathbb{S}^2)$ admit an orthonormal basis given by the *spherical harmonics (SH)* $\{Y_{\ell m}\}_{\ell \geq 0,\, |m| \leq \ell}$. Any $f \in L^2(\mathbb{S}^2)$ admits the expansion

$$f = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \mathcal{F}[f](\ell, m)\, Y_{\ell m}, \qquad \mathcal{F}[f](\ell, m) = \langle f, Y_{\ell m} \rangle_{L^2(\mathbb{S}^2)}.$$

The transformation $f \mapsto \mathcal{F}[f]$ is the *spherical harmonic transform (SHT)* (details in Appendix B).

On the sphere, convolution is defined via rotations rather than translations. Following Driscoll & Healy (1994), given functions $f, \kappa \in L^2(\mathbb{S}^2)$, their spherical convolution is defined as:

$$(f \star \kappa)(x) = \int_{\mathrm{SO}(3)} f(R\vec{n})\, \kappa(R^{-1}x)\, \mathrm{d}R, \quad \forall\, x \in \mathbb{S}^2, \tag{5}$$

where $\vec{n} = (0, 0, 1)^\top$ denotes the north pole and $\mathrm{d}R$ the (unnormalized) volume measure on $\mathrm{SO}(3)$.

The spherical convolution admits a convolution theorem.

Theorem 3.3 (Spherical Convolution Theorem, Driscoll & Healy, 1994)**.** *Let $f, \kappa \in L^2(\mathbb{S}^2)$. Then*

$$\forall l \geq 0, \forall |m| \leq \ell,\ \mathcal{F}[f \star \kappa](\ell, m) = 2\pi \sqrt{\frac{4\pi}{2\ell + 1}}\, \mathcal{F}[f](\ell, m)\, \mathcal{F}[\kappa](\ell, 0). \tag{6}$$

A key observation is that the heat kernel is *radial*, i.e. depends only on $\langle x, y \rangle$. For such kernels, the generic kernel operator (4) reduces to a spherical convolution (5). This contribution differs from prior work (Kondor & Trivedi, 2018; Sosnovik et al., 2021) in that it provides a direct link between the two convolution types, enabling the use of SHTs for efficient implementation of Sinkhorn iterations with the heat kernel. The proof is provided in Appendix A.3.

Proposition 3.4. *Let K be a radial kernel on $\mathbb{S}^2 \times \mathbb{S}^2$ of the form $K(x, y) = K_0(\langle x, y \rangle)$ with $K_0 \in L^2([-1, 1])$. Then for any $f \in L^2(\mathbb{S}^2)$, writing $\tilde{K}(z) = K(z, \vec{n})$ with $\vec{n}$ the north pole:*

$$\forall x \in \mathbb{S}^2, \quad (Kf)(x) = \frac{1}{2\pi}\big(f \star \tilde{K}\big)(x).$$

3.2 Algorithm

Each SH $Y_{\ell m}$ is an eigenfunction of the Laplace-Beltrami operator *i.e.* $-\Delta Y_{\ell m} = \ell(\ell+1)\, Y_{\ell m}$, so the heat operator acts diagonally: $\mathcal{F}[\mathcal{H}_t f](\ell, m) = e^{-t\ell(\ell+1)}\, \mathcal{F}[f](\ell, m)$. By Proposition 3.4 and the Spherical Convolution Theorem 3.3, applying the heat kernel to a function reduces to a SHT, pointwise multiplication by the precomputed weights $\{e^{-t\ell(\ell+1)}\}_\ell$, and an inverse SHT. We thus define the *heat convolution* as

$$\mathrm{HeatConv}(\mathbf{v};\, \{h_\ell\}) := \mathcal{F}^{-1}\big(\{h_\ell\} \odot \mathcal{F}(\mathbf{v})\big), \qquad h_\ell = e^{-t\ell(\ell+1)},$$

where $\odot$ denotes pointwise multiplication over SH coefficients. For numerical stability, classical stabilization techniques such as centering dual variables (Séjourné et al., 2023) can be applied. This yields Algorithm 1. Note that in order to match the standard Sinkhorn algorithm with regularization parameter ε, one must set the heat time to $t = \varepsilon/4$, for the theoretical reasons discussed in Section 3.1.

Discrete Setting and Complexity. In order to perform numerical computations, we discretize the sphere and represent probability distributions as vectors. We consider two discretizations: the HEALPix grid (Górski et al., 2005), which provides a hierarchical, equal-area partitioning of the sphere, and the equiangular Clenshaw-Curtis grid (Clenshaw & Curtis, 1960) (see Appendix C.1).

Given a discretization $\{x_i\}_{i=1}^n$ with area weights $\mathbf{a} \in \mathbb{R}_+^n$, we approximate integrals as $\int_{\mathbb{S}^2} f \,\mathrm{d}\sigma \approx \mathbf{a}^\top \mathbf{f}$ (see Appendix C.2). The usual Sinkhorn algorithm requires a memory cost of $\mathcal{O}(n^2)$ and a computational cost of $\mathcal{O}(n^2)$ per iteration. By contrast, our approach requires only $\mathcal{O}(n)$ memory for both grids, since no kernel matrix is ever formed. On the Clenshaw-Curtis grid with $n = \mathcal{O}(L^2)$ (where L is the spherical harmonic bandwidth), each SHT costs $\mathcal{O}(L^3) = \mathcal{O}(n^{3/2})$ operations; HEALPix with $n = 12N_{\mathrm{side}}^2$ (where N_{side} is the resolution parameter) achieves the same asymptotic cost via pseudo-spectral methods. In both cases, all operations, SHTs and pointwise multiplications, are dense and parallelizable, making the algorithm fully GPU-friendly. We provide further details on the practical implementation in Appendix C.

4 Experiments

4.1 Runtime comparison

Experimental setup. We benchmark all methods in the balanced case (*i.e.* $\tau = \infty$), as some baselines are restricted to this setting. We consider pairs of random probability measures on $\mathbb{S}^2$, generated by normalizing a uniform random field with respect to the area weights of the grid. Experiments are run on a single NVIDIA A40 GPU with `float64` precision. Each runtime is averaged over 5 independent runs. We consider both grids using powers of 2: $L \in \{8, \dots, 512\}$ for the equiangular (Clenshaw-Curtis) grid (up to $\approx$523,000 points) and $N_{\mathrm{side}} \in \{8, \dots, 256\}$ for HEALPix. We compare six methods. The first three use entropic regularization with $\varepsilon = 0.1$ and convergence threshold 10^{-9}: SHOT (ours), Sinkhorn (with matrix kernel convolutions), and Cholesky (Solomon et al., 2015) (convolutional approach with a pre-factored Cholesky decomposition of the Laplace-Beltrami operator, reimplemented in `PyTorch` (Paszke et al., 2019)). Two sliced methods using 500 projections: SSW (Bonet et al., 2023) and LSSOT (Liu et al., 2024). Finally, W_2^2 solves the exact OT linear program on pairwise geodesic distances via `POT` (Flamary et al., 2021). Note that for SHOT the restriction to the balanced case does not affect much the behavior of the method, as the only difference between the balanced and unbalanced formulations is an additional pointwise power step applied at each Sinkhorn iteration.

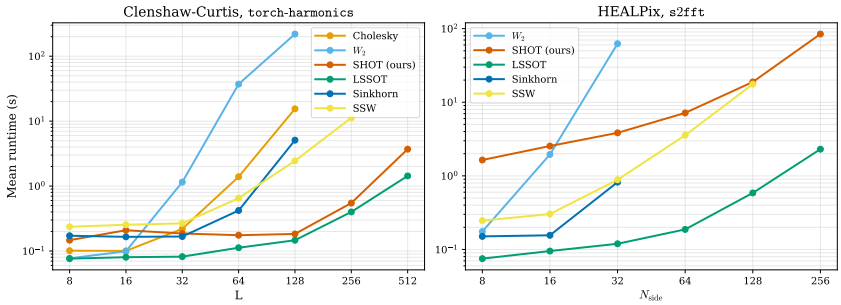


Figure 1: Mean runtime (seconds) vs. grid resolution averaged over 5 runs; missing values indicate out-of-memory. *Left*: Clenshaw-Curtis grid. *Right*: HEALPix grid.

Results. A key advantage of SHOT is its ability to scale to very high resolutions on both grids without running out of memory. As visible in Figure 1, methods relying on explicit kernel matrices (SINKHORN, CHOLESKY, W_2^2) hit out-of-memory errors at moderate resolutions, whereas SHOT successfully runs at the largest resolutions considered. On the Clenshaw-Curtis grid, SHOT is the second fastest method for $L \geq 64$, slightly behind LSSOT. On the HEALPix grid, however, SHOT does not achieve the same relative speedup, although its asymptotic complexity is identical to the Clenshaw-Curtis one: the gap is an implementation artifact rather than an algorithmic obstruction. Spherical convolutions on the Clenshaw-Curtis grid are performed via `torch-harmonics` (Bonev et al., 2023), which provides optimized CUDA kernels, whereas the HEALPix grid relies on `s2fft` (Price & McEwen, 2024), which does not. This limitation of the HEALPix spherical harmonic transform is acknowledged in Price & McEwen (2024, Section 4.2.4), and expected to be lifted by future developments of core `JAX` primitives. We provide more runtime comparisons in Appendix D.1, showing that this behavior is consistent across regularization parameters ε, grid types and in the unbalanced case.

4.2 EVALUATING CLIMATE MODELS ON THE SPHERE

4.2.1 SETUP AND MOTIVATION

We apply our SHOT method to compute the associated SD $\hat{S}_\varepsilon^\tau$ introduced in Section 3.1 to evaluate eight CMIP6 models against the ERA5 reanalysis (Hersbach et al., 2020), which we treat as the observational reference. In this study, we focus on precipitation fields which are represented as discrete measures on the HEALPix grid (Górski et al., 2005) at resolution $N_{\text{side}} = 64$ ($\sim$49 152 equal-area pixels, $\approx 1°$effective resolution), which provides an equal-area pixelation of $\mathbb{S}^2$ directly compatible with our algorithm. ERA5 data is accessed via the HERA5 dataset[1], already provided in HEALPix format, while CMIP6 model outputs are retrieved from the Pangeo Google Cloud archive[2] and regridded to the same HEALPix grid via bilinear interpolation. All fields are converted to mm/day. We analyze two climatological seasons: December-January-February (DJF) and June-July-August (JJA), each computed as a temporal average over the 2000–2014 historical period. Visual illustrations of the resulting distributions for ERA5 and one other model are shown in Appendix E.

Usefulness of UOT. CMIP6 models exhibit systematic global mean precipitation biases relative to ERA5, so the total mass of the two measures differs. A balanced transport plan is then forced to compensate for these intensity differences through spatial displacement, confounding the two sources of error. The SD $\hat{S}_\varepsilon^\tau$ yields a principled decomposition of model bias into a *spatial* component (transport of precipitation from wrong locations) and an *intensity* component (local over- or under-estimation of amounts), which is not achievable with any pointwise metric. The key advantage of SHOT is that it can efficiently compute $\hat{S}_\varepsilon^\tau$ at high resolution while other methods like CHOLESKY and SINKHORN require extensive computational resources to compute their associated SD at the considered resolution. We set $\varepsilon = 0.01$ (squared geodesic distance on the unit sphere) throughout; results are qualitatively stable over $\varepsilon \in [0.01, 1]$.

[1] `https://orcestra-campaign.org/hera5.html`
[2] `https://pangeo-data.github.io/pangeo-cmip6-cloud/`

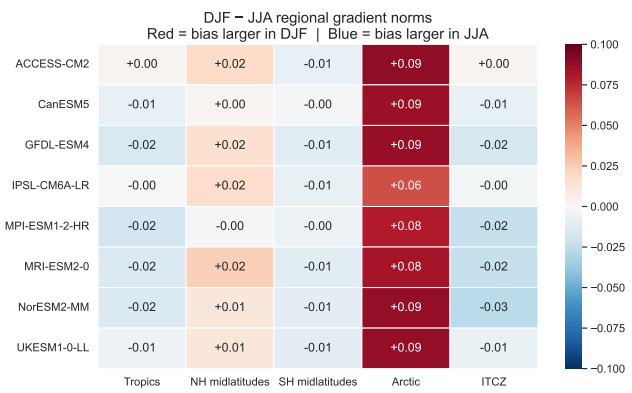


Figure 3: DJF−JJA difference of regional OT gradient norms under the unbalanced Sinkhorn divergence. Red: larger bias in DJF; blue: larger bias in JJA.

4.2.2 Global Model Ranking and Regional Bias Attribution

Scalar ranking. For each model m and season, we compute the balanced $\hat{S}_\varepsilon$ and unbalanced $\hat{S}_\varepsilon^\tau$ (with $\tau = 100$) Sinkhorn divergences between the precipitation measures μ_m and the reference μ_{ERA5}. The choice of τ was mostly guided by a numerical balancing of the magnitude of the KL divergences and the main OT cost and we provide a sensitivity analysis in Appendix E. In the balanced case, both measures are normalized to share the same total mass, while in the unbalanced case they are left unchanged. The resulting scalar score provides a geometrically meaningful ranking of models that accounts for both the spatial structure and, in the unbalanced case, the intensity of precipitation biases. Unlike RMSE, this score is a proper divergence as it equals zero if and only if $\mu_m = \mu_{\mathrm{ERA5}}$ a.e. and it metrizes weak convergence on $\mathbb{S}^2$, meaning that a model improving its score is genuinely moving its precipitation distribution closer to the reference in a transport sense.

Interestingly, rankings exhibit severe changes between the balanced and unbalanced cases. Figure 2 (panel 1) shows differences in those rankings for the two considered configurations and for all eight considered models in the DJF season. Large rank shifts reveal models whose scores under balanced OT are dominated by global intensity biases rather than spatial fidelity: GFDL-ESM4 improves from rank 5 to rank 1 (intensity bias masking good spatial structure), while NorESM2-MM degrades from rank 1 to rank 6 (balanced OT artificially rewarding compensating errors). MPI-ESM1-2-HR is the most metric-stable top performer (rank 2 in both). Figure 2 (panel 2) exhibits a seasonal comparison of unbalanced rankings (DJF vs. JJA). Four models have perfectly stable rankings across seasons, identifying them as the models with the most season-independent spatial fidelity and bias. Models improving in JJA (MRI-ESM2-0, IPSL-CM6A-LR) suggest boreal-summer specific improvements in spatial precipitation structure. We then turn to the question of *where* on the globe these improvements occur, which is not revealed by the scalar rankings but can be addressed by the gradient-based regional attribution described next.

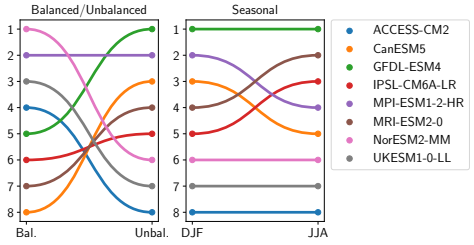


Figure 2: Model rankings from SD. *Left:* Balanced vs. unbalanced Sinkhorn divergences (DJF). *Right:* DJF vs. JJA unbalanced rankings.

Gradient-based regional attribution. The SD scalar score summarizes model quality but does not reveal *where* on the globe the bias originates. To address this, we exploit the fact that $\hat{S}_\varepsilon^\tau$ is differentiable with respect to μ_m. Its gradient $g_m = \nabla_{\mu_m} \hat{S}_\varepsilon^\tau(\mu_m, \mu_{\mathrm{ERA5}})$ is a signed field on $\mathbb{S}^2$, derived from the Kantorovich potentials of the dual problem, where $g_m(x) > 0$ indicates local excess precipitation mass and $g_m(x) < 0$ indicates deficit relative to the optimal transport plan. Since Kantorovich potentials are defined up to an additive constant, we center g_m by subtracting its area-weighted mean, retaining only the spatially varying component.

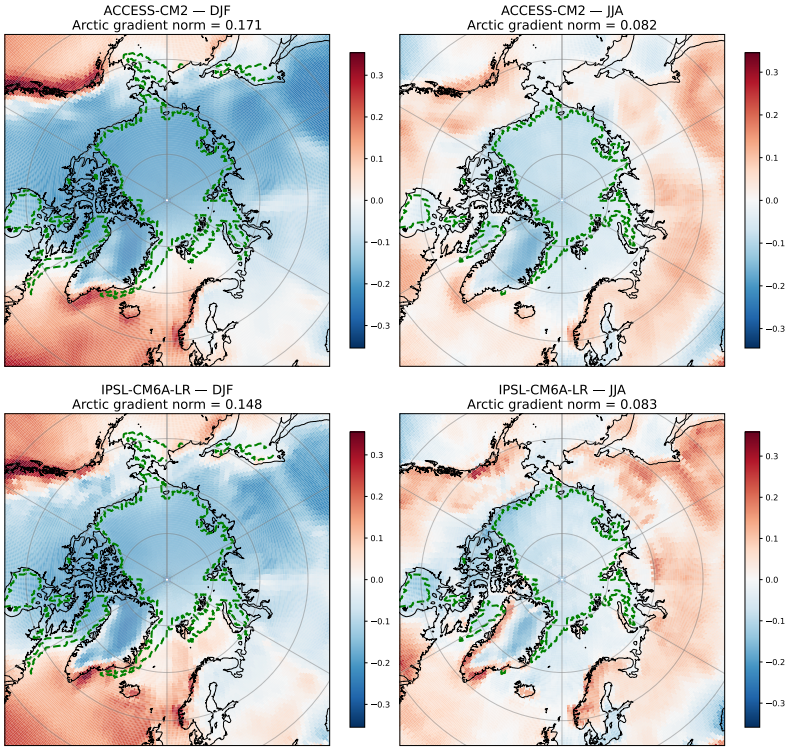


Figure 4: OT gradient of the Sinkhorn divergence (centered) over the Arctic for ACCESS-CM2 (top) and IPSL-CM6A-LR (bottom), in DJF (left) and JJA (right). Green contours show the ERA5 climatological sea-ice edge at 15% concentration for DJF and JJA. Red: model has excess precipitation; blue: deficit. Arctic gradient norms are reported in the panel titles.

We partition $\mathbb{S}^2$ into five non-overlapping geographic regions: Tropics ($|\text{lat}| < 23.5°$), ITCZ ($|\text{lat}| < 10°$), NH Midlatitudes ($30° - 60°$N), SH Midlatitudes ($30° - 60°$S), and Arctic ($> 66.5°$N), and compute the area-weighted RMS of g_m over each region as a regional bias index. We provide in Appendix E, Figure 13 these indices for all eight models and both seasons, normalized by the global maximum to preserve inter-model differences. We now precisely focus on the seasonal contrast (*i.e.* difference, denoted by Δ in the following) of these regional indices, which is not visible in the scalar rankings but reveals physically interpretable patterns of model bias and improvement. The Arctic column in Figure 3 is the *only* region with a meaningful seasonal signal, showing consistently larger DJF bias across all eight models ($\Delta = +0.06$ to $+0.09$). All other regions show differences of $|\Delta| \leq 0.02$, confirming that tropical, midlatitude, and ITCZ precipitation biases are structurally season-independent. The slight inter-model variation in Arctic seasonal contrast (IPSL-CM6A-LR at $+0.06$ versus most others at $+0.09$) is physically interpretable as reflecting differences in sea-ice parameterization complexity, as discussed below.

4.2.3 Arctic Bias and Sea-Ice Coupling

The seasonal collapse of the Arctic gradient norm, from a strongly discriminating signal in DJF to near-zero in JJA across all models, points to a specific physical mechanism: Arctic precipitation in boreal winter is tightly coupled to sea-ice dynamics, through moisture fluxes over the marginal ice zone, polar vortex interactions, and storm track activity (Serreze & Meier, 2019). We investigate this coupling directly by overlaying the ERA5 climatological sea-ice edge (15% concentration contour) on the gradient maps of the two most contrasting models: ACCESS-CM2 (largest Arctic seasonal contrast, $\Delta = 0.089$) and IPSL-CM6A-LR (smallest contrast, $\Delta = 0.065$). Figure 4 shows the centered gradient g_m over the Arctic in NorthPolarStereo projection for both models and both seasons, with the DJF and JJA sea-ice edges overlaid.

Results. Three findings stand out. First, in DJF the gradient signal is *spatially organized around the sea-ice edge* in both models: a precipitation deficit appears inside the ice pack and an excess appears along the marginal ice zone and just outside it. This dipole structure is consistent with a systematic equatorward displacement or overextension of the modeled ice edge, which shifts the precipitation maximum outward relative to ERA5. The OT gradient localizes this bias precisely, which a difference map cannot, since the latter conflates the displacement with a local intensity error.

Second, the two models exhibit qualitatively different bias geometries in DJF. ACCESS-CM2 shows a large-scale, coherent dipole spanning the entire Arctic basin (deficit interior, excess periphery), suggesting a bulk ice-edge displacement bias. IPSL-CM6A-LR shows a more fragmented, spatially heterogeneous pattern concentrated along the marginal ice zone, consistent with a fine-scale positional error rather than a global shift. This distinction, directly readable from the gradient geometry, is not recoverable from any scalar metric.

Third, in JJA both models converge to similar gradient patterns (norms 0.082 and 0.083 respectively) as the ice retreats and its organizing influence weakens. The residual JJA signal is predominantly positive (model overestimates precipitation over the open Arctic Ocean in summer), consistent with known warm biases in Arctic sea-surface temperatures in coupled models. The near-identical JJA patterns confirm that the ACCESS-CM2 vs. IPSL-CM6A-LR difference is specifically a boreal winter phenomenon driven by sea-ice representation, not a general Arctic precipitation deficiency.

Summary. The experiments demonstrate that spherical unbalanced Sinkhorn divergences can provide a principled, geometrically faithful, and physically interpretable framework for climate model evaluation. Beyond scalar rankings, the differentiability of the divergence yields gradient maps that attribute model bias to specific geographic regions and, in the Arctic, reveal a possible mechanistic link to sea-ice edge geometry that is invisible to standard metrics.

5 Discussion and Future Work

In this work, we introduced SHOT, a novel algorithm for computing the Sinkhorn divergence associated with the heat kernel on the sphere. We demonstrated its efficiency and applicability to high-resolution data, and illustrated its potential for evaluating climate models in a geometry-aware manner. However, several limitations and avenues for future work remain. First, as every entropy regularized OT method, SHOT is subject to numerical instabilities for small ε, which limits its ability to approximate the true OT cost, as discussed in in Appendix C.4. To some extent, this can be amplified by the spiking behavior of the heat kernel, which can produce negative values when truncated to a finite degree. Developing strategies to mitigate these instabilities and extend the operating range of the algorithm is an important direction for future work. One potential remedy is to combine the heat kernel with suitable filters to dampen its spiking behavior. On the theoretical side, the framework is not specific to the sphere: Theorem 3.1 holds on any compact connected Riemannian manifold and Proposition 3.2 on compact connected symmetric spaces, a class containing flat tori and compact Lie groups such as $\mathrm{SO}(3)$ and $\mathrm{SU}(2)$. The restriction is algorithmic: fast harmonic transforms are mature only for a few canonical examples (Gräf & Hielscher, 2015; Delgado & Umaña, 2026), and the approach should extend to these spaces as such transforms become available. Finally, since $\hat{S}_\varepsilon^\tau$ is fully differentiable, it could serve as a geometry-aware training objective for networks operating on spherical data. Two directions seem particularly promising: as a training or fine-tuning loss for spherical neural operators such as SFNO-based weather emulators (Bonev et al., 2023), which are known to blur and lose spectral sharpness under pointwise losses, and for which the memory and speed of SHOT at operational resolutions is precisely what is required; and as a reconstruction loss for implicit neural representations of spherical fields (Hanon et al., 2026), which still rely on pointwise objectives despite their intrinsically non-Euclidean domain. Assessing the benefit of OT losses in these settings is a question on its own, which we leave to future work. Lastly and regarding climate model evaluation, our goal was to illustrate the potential of the method rather than to provide a comprehensive assessment of CMIP6 models, which would require a broader analysis spanning multiple variables, seasons, and historical periods. Future work will explore these dimensions and investigate the use of the method for model development and tuning.

AI USE STATEMENT

Generative AI tools were used only to polish the writing of this paper: rephrasing and proofreading sentences that had already been written by the authors. They were not used for research ideation, for deriving or checking any of the theoretical results, for writing the code, or for producing or analysing the experimental results. All AI-assisted text was reviewed by the authors, and we take responsibility for the final content of this work.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 214–223. PMLR, July 2017.

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.

Clément Bonet, Kimia Nadjahi, Thibault Séjourné, Kilian Fatras, and Nicolas Courty. Slicing Unbalanced Optimal Transport. *Transactions on Machine Learning Research*, 2024.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical Fourier neural operators: Learning stable dynamics on the sphere. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2806–2823. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/bonev23a.html`.

Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

Maciej Buze and Manh Hong Duong. Entropic regularisation of unbalanced optimal transportation problems. *arXiv preprint arXiv:2305.02410*, 2023.

Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2): 1385–1418, 2017.

Guillaume Carlier, Paul Pegon, and Luca Tamanini. Convergence rate of general entropic optimal transport costs. *Calculus of Variations and Partial Differential Equations*, 62(4):116, 2023.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303.

C. W. Clenshaw and A. R. Curtis. A method for numerical integration on an automatic computer. *Numerische Mathematik*, 2:197–205, 1960.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9): 1853–1865, September 2017. ISSN 1939-3539. doi: 10.1109/TPAMI.2016.2615921.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pp. 2292–2300. Curran Associates, Inc., 2013.

Nathaël Da Costa, Cyrus Mostajeran, Juan-Pablo Ortega, and Salem Said. Invariant kernels on riemannian symmetric spaces: a harmonic-analytic approach. *SIAM Journal on Mathematics of Data Science*, 7(2):752–776, 2025.

Julio Delgado and Alejandro Umaña. On the Fast Fourier Transform on SU(2), 2026. URL https://arxiv.org/abs/2605.23923.

Kathryn P Drake and Grady B Wright. A fast and accurate algorithm for spherical harmonic analysis on healpix grids with applications to the cosmic microwave background radiation. *Journal of Computational Physics*, 416:109544, 2020.

J. R. Driscoll and D. M. Healy. Computing Fourier transforms and convolutions on the 2-sphere. *Advances in applied mathematics*, 15(2):202–250, 1994.

Jost-Hinrich Eschenburg. Lecture notes on symmetric spaces. *preprint*, 1997.

V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

J. Faraut. Analysis on lie groups. *Cambridge studies in advanced mathematics*, 110, 2008.

Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3032–3042, 2015.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690. PMLR, 16–18 Apr 2019.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Sylvain Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021.

Robert Garrett, Trevor Harris, Zhuo Wang, and Bo Li. Validating climate models with spherical convolutional wasserstein distance. *Advances in Neural Information Processing Systems*, 37: 59119–59149, 2024.

Krzysztof M. Górski, Eric Hivon, A. J. Banday, B. D. Wandelt, Frode K. Hansen, Mstvos Reinecke, and Matthias Bartelmann. HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759–771, 2005.

Manuel Gräf and Ralf Hielscher. Fast global optimization on the torus, the sphere, and the rotation group. *SIAM Journal on Optimization*, 25(1):540–563, 2015. doi: 10.1137/130950070. URL https://doi.org/10.1137/130950070.

Alexander Grigoryan. *Heat kernel and analysis on manifolds*, volume 47. American Mathematical Soc., 2009.

Théo Hanon, Nicolas Mil-Homens Cavaco, John Kiely, and Laurent Jacques. Herglotz-net: Implicit neural representation of spherical data with harmonic positional encoding, 2026. URL https://arxiv.org/abs/2502.13777.

Dennis M Healy Jr, Daniel N Rockmore, Peter J Kostelec, and Sean Moore. Ffts for the 2-sphere-improvements and variations. *Journal of Fourier analysis and applications*, 9(4):341–385, 2003.

Sigurdur Helgason. *Differential geometry, Lie groups, and symmetric spaces*, volume 80. Academic press, 1979.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Sangwon Hyun, Aditya Mishra, Christopher L. Follett, Bror Jonsson, Gemma Kulk, Gael Forget, Marie-Fanny Racault, Thomas Jackson, Stephanie Dutkiewicz, Christian L. Müller, and Jacob Bien. Ocean mover's distance: using optimal transport for analysing oceanographic data. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 478(2262), 2022.

Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. Kernel methods on riemannian manifolds with gaussian rbf kernels. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2464–2477, 2015.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International conference on machine learning*, pp. 2747–2755. PMLR, 2018.

Stefan Kunis and Daniel Potts. Fast spherical Fourier algorithms. *Journal of Computational and Applied Mathematics*, 161(1):75–98, 2003.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3): 969–1117, 2018.

Xinran Liu, Yikun Bai, Rocio Diaz Martin, Kaiwen Shi, Ashkan Shahbazi, Bennett Allan Landman, Catie Chang, and Soheil Kolouri. Linear Spherical Sliced Optimal Transport: A Fast Metric for Comparing Spherical Data. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley, 2 edition, 2000.

Jason D McEwen and Yves Wiaux. A novel sampling theorem on the sphere. *IEEE Transactions on Signal Processing*, 59(12):5876–5887, 2011.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Luca Nenna, Paul Pegon, and Louis Tocquec. Convergence rates for regularized unbalanced optimal transport: the discrete case. *arXiv preprint arXiv:2507.07917*, 2025.

Marcel Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Edward Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

Matthew A Price and Jason D McEwen. Differentiable and accelerated spherical harmonic and wigner transforms. *Journal of Computational Physics*, 510:113109, 2024.

Michael Quellmalz, Robert Beinert, and Gabriele Steidl. Sliced optimal transport on the sphere. *Inverse Problems*, 39(10):105005, August 2023. ISSN 0266-5611. doi: 10.1088/1361-6420/acf156.

Michael Quellmalz, Léo Buecher, and Gabriele Steidl. Parallelly sliced optimal transport on spheres and on the rotation group. *Journal of Mathematical Imaging and Vision*, 66(6):951–976, 2024.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87. Birkhäuser, 2015.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Mark C. Serreze and Walter N. Meier. The Arctic's sea ice cover: trends, variability, predictability, and comparisons to the Antarctic. *Annals of the New York Academy of Sciences*, 1436(1):36–53, January 2019. doi: 10.1111/nyas.13856. Epub 2018 May 28.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

Ivan Sosnovik, Artem Moskalev, and Arnold Smeulders. Disco: accurate discrete scale convolutions. *arXiv preprint arXiv:2106.02733*, 2021.

Franziskus Steinert, Salem Said, and Cyrus Mostajeran. Universal kernels via harmonic analysis on riemannian symmetric spaces. In *International Conference on Geometric Science of Information*, pp. 172–180. Springer, 2025.

Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. In Emmanuel Trélat and Enrique Zuazua (eds.), *Numerical Control: Part B*, volume 24 of *Handbook of Numerical Analysis*, pp. 407–471. Elsevier, 2023. doi: https://doi.org/10.1016/bs.hna.2022.11.003. URL https://www.sciencedirect.com/science/article/pii/S1570865922000230.

Huy Tran, Yikun Bai, Abihith Kothapalli, Ashkan Shahbazi, Xinran Liu, Rocio P Diaz Martin, and Soheil Kolouri. Stereographic spherical sliced Wasserstein distances, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tran24a.html.

S. R. S. Varadhan. On the behavior of the fundamental solution of the heat equation with variable coefficients. *Communications on Pure and Applied Mathematics*, 20(2):431–455, 1967.

Carlos Veiga Rodrigues and Io Odderskov. Climate error metrics based on wasserstein distances. *Applied Energy*, 398, 2025.

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

Gabriele Vissio, Valerio Lembo, Valerio Lucarini, and Michael Ghil. Evaluating the performance of climate models based on wasserstein distance. *Geophysical Research Letters*, 47(21), 2020.

Zhiang Xie, Dongwei Chen, and Puxi Li. Discovering climate change during the early 21st century via wasserstein stability analysis. *Advances in Atmospheric Sciences*, 42:373–381, 02 2025.

Felix X-F Ye, Xingjie Li, An Yu, Ming-Ching Chang, Linsong Chu, and Davis Wertheimer. Flashsinkhorn: Io-aware entropic optimal transport. *arXiv preprint arXiv:2602.03067*, 2026.

A Proofs

A.1 Proof of Theorem 3.1

We will need some preliminaries. First, note that by Varadhan's celebrated short-time asymptotics of the heat kernel, one has uniform convergence of the time-ε cost to the squared geodesic distance (Varadhan, 1967, Theorem 2.2) on the compact, connected Riemannian manifold $\mathcal{M}$:

$$\widehat{C}_\varepsilon(x,y) = -\varepsilon \log \mathcal{H}_{\frac{\varepsilon}{4}}(x,y) \xrightarrow[\varepsilon\to 0]{\text{uniformly}} d(x,y)^2.$$

Second, we will need some properties of the KL divergence, notably that it is jointly lower semi-continuous (see e.g. Posner (1975, Theorem 1)) and that it is superlinear. For convenience of the reader, we state and reprove that latter statement.

Lemma A.1 (Superlinear lower bound for generalized KL)**.** *Let $\nu \in \mathcal{R}_+(\mathcal{M})$ with $\nu(\mathcal{M}) > 0$. Then for every $\mu \in \mathcal{R}_+(\mathcal{M})$ with $\mu \ll \nu$,*

$$\mathrm{KL}(\mu \mid \nu) \;\geq\; \mu(\mathcal{M}) \log \frac{\mu(\mathcal{M})}{\nu(\mathcal{M})} - \mu(\mathcal{M}) + \nu(\mathcal{M}).$$

In particular, the right-hand side is superlinear in $\mu(\mathcal{M})$.

Proof. Applying Jensen's inequality to the convex function $t \mapsto t \log t$ with respect to the probability measure $\nu/\nu(\mathcal{M})$ gives

$$\mathrm{KL}(\mu \mid \nu) = \int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \,\mathrm{d}\nu - \mu(\mathcal{M}) + \nu(\mathcal{M}) \geq \mu(\mathcal{M}) \log \frac{\mu(\mathcal{M})}{\nu(\mathcal{M})} - \mu(\mathcal{M}) + \nu(\mathcal{M}). \quad \square$$

We will also use the fact that the problem $\mathrm{UOT}^{\tau}_{d^2}$ admits a minimizer. This is a consequence of Liero et al. (2018, Theorem 3.3) as the conditions of that theorem are satisfied: the problem is *feasible* (take the null transport plan) and the KL divergence is superlinear (as stated just above).

To the best of our knowledge, no full proof of a similar result exists for entropic regularized UOT: Buze & Duong (2023, Theorem 3.3) announce such a result for a future version of their paper but currently none is present while Nenna et al. (2025, Proposition 2.3) prove existence and uniqueness for a slightly more general context but restrict themselves to the discrete case, using arguments similar to those of the balanced case i.e. essentially the direct method of the calculus of variations and strict convexity of the regularizer.

As we will need such a result to study $\mathrm{UOT}^{\tau}_{\widehat{C}_\varepsilon,\varepsilon}$, before proceeding, we remark here that arguing along the same lines is sufficient to establish existence and uniqueness beyond the discrete case : observe first that by lower semicontinuity of the KL divergence, the functional defining regularized UOT is lower semicontinuous. Moreover, for $\varepsilon > 0$, superlinearity of the KL divergence implies that a minimizing sequence must be bounded in mass (see e.g. Liero et al. (2018, Estimate 3.10)) and thus, by Prokhorov's theorem for finite positive measures applied to the compact space $\mathcal{M}^2$, admits a converging subsequence in the topology of weak convergence of measures (this is essentially stating that the associated functional is *coercive*) and existence of the minimizer follows. That the minimizer is unique derives from the strict convexity of the KL divergence regularizer.

Finally, we need the following technical lemma (inspired by Nutz (2021, Lemma 5.3)).

Lemma A.2 (Block approximation)**.** *Let X be a compact metric space, let $P, Q \in \mathcal{R}_+(X)$, and let $\pi \in \mathcal{R}_+(X^2)$ be such that $\mathrm{KL}(\pi^1 \mid P) < +\infty$ and $\mathrm{KL}(\pi^2 \mid Q) < +\infty$. Let moreover $f : X^2 \to \mathbb{R}$ be a continuous function. Then for every $\eta > 0$, there exists $\pi_\eta \in \mathcal{R}_+(X^2)$ such that $\mathrm{KL}(\pi_\eta \mid P \otimes Q) < +\infty$, having the same mass as π and satisfying*

1. *$\mathrm{KL}(\pi^1_\eta \mid P) \leq \mathrm{KL}(\pi^1 \mid P)$ and $\mathrm{KL}(\pi^2_\eta \mid Q) \leq \mathrm{KL}(\pi^2 \mid Q)$,*
2. *$\left|\int_{X^2} f \, d\pi - \int_{X^2} f \,\mathrm{d}\pi_\eta\right| \leq \eta \, m(\pi)$.*

Moreover, if $m(P) = m(Q)$ and π is a coupling between P and Q, then π_η is also a coupling between P and Q.

Proof. Let $\eta > 0$. Since f is continuous and X^2 is compact, f is uniformly continuous. Hence there exists $\delta > 0$ such that $d(x, x') \leq \delta,\ d(y, y') \leq \delta \implies |f(x, y) - f(x', y')| \leq \eta$.

By compactness of X, there exists a finite measurable partition $\{A_1, \ldots, A_N\}$ of X such that $\mathrm{diam}(A_i) \leq \delta$ for all i. Define $I^+ := \{(i, j) : P(A_i) > 0,\ Q(A_j) > 0\}$ and set

$$\pi_\eta := \sum_{(i,j) \in I^+} \frac{\pi(A_i \times A_j)}{P(A_i) Q(A_j)} (P \otimes Q)|_{A_i \times A_j}.$$

By construction $\pi_\eta \ll P \otimes Q$ and $\mathrm{KL}(\pi_\eta \mid P \otimes Q) < +\infty$ and moreover

$$\frac{d\pi_\eta}{d(P \otimes Q)} = \sum_{(i,j)\in I^+} \frac{\pi(A_i \times A_j)}{P(A_i)Q(A_j)} \mathbf{1}_{A_i \times A_j}.$$

We now study the marginals. Since $\mathrm{KL}(\pi^1 \mid P) < +\infty$ we have $\pi^1 \ll P$, thus the Radon-Nikodym derivative $\frac{d\pi^1}{dP}$ exists and for every measurable $B \subset X$, one has

$$\pi^1_\eta(B) = \pi_\eta(B \times X) = \sum_{(i,j)\in I^+} \frac{\pi(A_i \times A_j)}{P(A_i)Q(A_j)} P(B \cap A_i)\, Q(A_j) = \sum_i \frac{\pi^1(A_i)}{P(A_i)} P(B \cap A_i),$$

as $\{A_j\}_j$ is a partition of X. Therefore,

$$\pi^1_\eta = \sum_i \frac{\pi^1(A_i)}{P(A_i)} P|_{A_i},$$

and thus

$$\frac{d\pi^1_\eta}{dP} = \sum_i \left(\frac{1}{P(A_i)} \int_{A_i} \frac{d\pi^1}{dP}\, \mathrm{d}P \right) \mathbf{1}_{A_i}.$$

Recalling that

$$\mathrm{KL}(\mu \mid \nu) = \int \varphi\left(\frac{d\mu}{d\nu}\right) \mathrm{d}\nu \text{ with } \varphi(t) := t \log t - t + 1,$$

the convexity of φ and Jensen's inequality in normalized form yield

$$\varphi\left(\frac{1}{P(A_i)} \int_{A_i} \frac{d\pi^1}{dP}\, \mathrm{d}P \right) \le \frac{1}{P(A_i)} \int_{A_i} \varphi\left(\frac{d\pi^1}{dP}\right) \mathrm{d}P.$$

Multiplying by $P(A_i)$ and summing over i, we obtain

$$\mathrm{KL}(\pi^1_\eta \mid P) = \sum_i P(A_i) \varphi\left(\frac{1}{P(A_i)} \int_{A_i} \frac{d\pi^1}{dP}\, \mathrm{d}P \right) \le \sum_i \int_{A_i} \varphi\left(\frac{d\pi^1}{dP}\right) \mathrm{d}P = \mathrm{KL}(\pi^1 \mid P).$$

This proves the first part of (1). The proof of the second part is identical.

We now prove that π and π_η have the same mass. Observe that if $(i,j) \notin I^+$, then

$$P(A_i) = 0 \qquad \text{or} \qquad Q(A_j) = 0.$$

If $P(A_i) = 0$, then since $\pi^1 \ll P$,

$$\pi(A_i \times A_j) \le \pi(A_i \times X) = \pi^1(A_i) = 0,$$

as $\pi^1 \ll P$. Similarly, if $Q(A_j) = 0$, then $\pi(A_i \times A_j) = 0$ $(i,j) \notin I^+$. Therefore,

$$m(\pi_\eta) = \sum_{(i,j)\in I^+} \frac{\pi(A_i \times A_j)}{P(A_i)Q(A_j)} (P \otimes Q)(A_i \times A_j) = \sum_{(i,j)\in I^+} \pi(A_i \times A_j) = \sum_{i,j} \pi(A_i \times A_j) = m(\pi).$$

Finally, we prove (2). From what precedes, as $\pi(A_i \times A_j) = 0$ for all $(i,j) \notin I^+$, one has

$$\int_{X^2} f\, \mathrm{d}\pi = \sum_{(i,j)\in I^+} \int_{A_i \times A_j} \int f \mathrm{d}\pi.$$

Define

$$\bar{f}_{ij} := \frac{1}{P(A_i)Q(A_j)} \int_{A_i \times A_j} f\, \mathrm{d}(P \otimes Q).$$

Then

$$\int_{X^2} f\, \mathrm{d}\pi_\eta = \sum_{(i,j)\in I^+} \pi(A_i \times A_j) \bar{f}_{ij},$$

hence

$$\int_{X^2} f \,\mathrm{d}\pi - \int_{X^2} f \,\mathrm{d}\pi_\eta = \sum_{(i,j)\in I^+} \int_{A_i\times A_j} (f - \bar{f}_{ij}) \,\mathrm{d}\pi.$$

Fix $(i,j) \in I^+$ and $(x,y) \in A_i \times A_j$. Then

$$\begin{aligned} |f(x,y) - \bar{f}_{ij}| &= \left| \frac{1}{P(A_i)Q(A_j)} \int_{A_i\times A_j} (f(x,y) - f(x',y')) \,\mathrm{d}(P\otimes Q)(x',y') \right| \\ &\le \frac{1}{P(A_i)Q(A_j)} \int_{A_i\times A_j} |f(x,y) - f(x',y')| \,\mathrm{d}(P\otimes Q)(x',y'). \end{aligned}$$

Since the A_i have diameter less than δ and $(x,y) \in A_i \times A_j$, the uniform continuity property mentioned at the beginning gives, for all $(x,y) \in A_i \times A_j$, $|f(x,y) - \bar{f}_{ij}| \le \eta$. Consequently,

$$\left| \int_{X^2} f \,\mathrm{d}\pi - \int_{X^2} f \,\mathrm{d}\pi_\eta \right| \le \sum_{(i,j)\in I^+} \int_{A_i\times A_j} |f - \bar{f}_{ij}| \,\mathrm{d}\pi \le \eta \sum_{(i,j)\in I^+} \pi(A_i \times A_j) = \eta\, m(\pi).$$

Assume now that $m(P) = m(Q)$ and that π is a coupling between P and Q, that is $\pi^1 = P$ and $\pi^2 = Q$. Then $\frac{\pi^1(A_i)}{P(A_i)} = 1$ thus $\pi^1_\eta = \sum_i P|_{A_i} = P$, since $\{A_i\}$ is a partition of X. Similarly, $\pi^2_\eta = Q$ and therefore π_η is also a coupling between P and Q. □

Proof of Theorem 3.1. We treat first the case $\tau < \infty$. The case $\tau = +\infty$ can be treated by similar arguments, with minor modifications, and is discussed at the end of the proof.

Case 1: $\tau < \infty$**.** On the space $\mathcal{R}_+(\mathcal{M}^2)$, consider the functional associated with the $\mathrm{UOT}^\tau_{d^2}$ problem:

$$F(\pi) = \int_{\mathcal{M}^2} d^2 \mathrm{d}\pi + \tau \left(\mathrm{KL}(\pi^1 \mid P) + \mathrm{KL}(\pi^2 \mid Q) \right),$$

and that associated with the heat kernel regularized $\mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}$, formally defined for $\varepsilon > 0$ by:

$$\hat{F}_\varepsilon(\pi) = \int_{\mathcal{M}^2} \widehat{C}_\varepsilon \,\mathrm{d}\pi + \tau \left(\mathrm{KL}(\pi^1 \mid P) + \mathrm{KL}(\pi^2 \mid Q) \right) + \varepsilon \,\mathrm{KL}(\pi \mid P \otimes Q).$$

By construction, $\mathrm{UOT}^\tau_{d^2}(P,Q)$ is the infimum of F while the heat-regularized cost $\mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}(P,Q)$ is the infimum value of F_ε. Let $\pi^* \in \mathcal{R}_+(\mathcal{M}^2)$ be a minimizer of F^τ *i.e.* $F^\tau(\pi^*) = \mathrm{UOT}^\tau_{d^2}(P,Q)$.

The proof proceeds in two steps: a limsup inequality and a liminf inequality.

Step 1: Upper bound. We show that

$$\limsup_{\varepsilon\to 0} \ \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}(P,Q) \le \mathrm{UOT}^\tau_{d^2}(P,Q) = F(\pi^*).$$

Note that π^* may have $\mathrm{KL}(\pi^* \mid P\otimes Q) = +\infty$; this is why we cannot simply take π^* as a competitor for F^τ_ε. Instead, we use the block approximation lemma A.2 with $X = \mathcal{M}, \pi = \pi^\star$ and $f = d^2$. Fixing $\eta > 0$, and letting π_η be given by that lemma, one has by definition,

$$\begin{aligned} \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}(P,Q) &\le F^\tau_\varepsilon(\pi_\eta) \\ &\le \int \widehat{C}_\varepsilon \,\mathrm{d}\pi_\eta + \varepsilon \mathrm{KL}(\pi_\eta \mid P\otimes Q) + \tau\big(\mathrm{KL}((\pi^*)^1 \mid P) + \mathrm{KL}((\pi^*)^2 \mid Q)\big), \end{aligned}$$

where the inequality comes from item (1) of the Lemma A.2. Write the difference in the transport cost:

$$\begin{aligned} \left| \int \widehat{C}_\varepsilon \,\mathrm{d}\pi_\eta - \int d^2 \,\mathrm{d}\pi^* \right| &\le \left| \int (\widehat{C}_\varepsilon - d^2) \,\mathrm{d}\pi_\eta \right| + \left| \int d^2 \,\mathrm{d}\pi_\eta - \int d^2 \,\mathrm{d}\pi^* \right| \\ &\le \|\widehat{C}_\varepsilon - d^2\|_\infty m(\pi_\eta) + \eta\, m(\pi^\star) = (\|\widehat{C}_\varepsilon - d^2\|_\infty + \eta)\, m(\pi^\star), \end{aligned}$$

by item (2) of the Lemma A.2 and the fact that π_η and $\pi^\star$ have the same mass. Consequently, one has,

$$\begin{aligned}\mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}(P,Q) \leq & (\|\widehat{C}_\varepsilon - d^2\|_\infty + \eta)\, m(\pi^\star) \\ & + \int d^2 \,\mathrm{d}\pi^* + \varepsilon \mathrm{KL}(\pi_\eta \mid P \otimes Q) + \tau\big(\mathrm{KL}((\pi^\star)^1 \mid P) + \mathrm{KL}((\pi^\star)^2 \mid Q)\big).\end{aligned}$$

As η is fixed independently of ε, we can let $\varepsilon \to 0$ and get, by Varadhan's formula

$$\begin{aligned}\limsup_{\varepsilon\to 0}\ \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}(P,Q) &\leq \eta m(\pi^\star) + \int d^2 \,\mathrm{d}\pi^* + \tau\big(\mathrm{KL}((\pi^\star)^1 \mid P) + \mathrm{KL}((\pi^\star)^2 \mid Q)\big) \\ &= \eta m(\pi^\star) + \mathrm{UOT}^\tau_{d^2}(P,Q).\end{aligned}$$

As $\eta > 0$ was arbitrary, we obtain the claimed upper bound.

Step 2: Lower bound. We show that

$$\mathrm{UOT}^\tau_{d^2}(P,Q) = F(\pi^*) \leq \liminf_{\varepsilon\to 0} \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}(P,Q).$$

For each $\varepsilon > 0$, let $\pi_\varepsilon \in \mathcal{R}_+(\mathcal{M}^2)$ be a minimizer of $\hat{F}_\varepsilon$.

From the first step, one sees that the limsup of the family $\{\mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}(P,Q)\}_\varepsilon$ is finite and as a consequence, that family is bounded, say by $\kappa > 0$.

By Varadhan's uniform convergence $\widehat{C}_\varepsilon \to d^2 \geq 0$ on the compact manifold $\mathcal{M}$, there exists $\varepsilon_0 > 0$ such that $\widehat{C}_\varepsilon \geq -1$ for all $\varepsilon \leq \varepsilon_0$. Since $\mathrm{KL}(\pi^2_\varepsilon \mid Q) \geq 0$ and $\mathrm{KL}(\pi_\varepsilon \mid P \otimes Q) \geq 0$, we may drop these terms and apply Lemma A.1 to the first marginal to obtain

$$\tau\Big(m(\pi^1_\varepsilon)\log\frac{m(\pi^1_\varepsilon)}{b} - m(\pi^1_\varepsilon) + b\Big) \ \leq\ \tau\,\mathrm{KL}(\pi^1_\varepsilon \mid P) \ \leq\ \kappa + m(\pi^1_\varepsilon).$$

Since the left-hand side is superlinear and the right-hand side is affine in $m(\pi^1_\varepsilon)$, we deduce that $m(\pi_\varepsilon) = m(\pi^1_\varepsilon)$ is bounded uniformly for $\varepsilon \leq \varepsilon_0$ i.e. the family $(\pi_\varepsilon)_{\varepsilon\leq\varepsilon_0}$ is bounded in mass.

By definition of the $\liminf$, there exists a sequence $(\varepsilon_k)_k$ converging to 0 such that

$$\mathrm{UOT}^\tau_{\widehat{C}_{\varepsilon_k},\varepsilon_k}(P,Q) \longrightarrow \liminf_{\varepsilon\to 0} \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon,\varepsilon}(P,Q).$$

We let k_0 be such that $\varepsilon_k \leq \varepsilon_0$ for $k \geq k_0$ and henceforth work with the sequence $(\pi_k)_{k\geq k_0}$. To simplify notation, we write $\widehat{C}_{\varepsilon_k} := \widehat{C}_k$, $F_{\varepsilon_k} = F_k$ and $\pi_k := \pi_{\varepsilon_k}$.

We now apply Prokhorov's theorem on $\mathcal{R}_+(\mathcal{M}^2)$: as $\mathcal{M}^2$ is compact and $(\pi_k)_{k\geq k_0}$ is bounded in mass, the sequence $(\pi_k)_{k\geq k_0}$ is relatively compact in $\mathcal{R}_+(\mathcal{M}^2)$. Extract a subsequence π_{k_n} that converges weakly to some $\pi \in \mathcal{R}_+(\mathcal{M}^2)$.

By definition of $\pi^\star$ we have

$$\mathrm{UOT}^\tau_{d^2}(P,Q) = F(\pi^\star) \leq F(\pi) = \int_{\mathcal{M}^2} d^2 \mathrm{d}\pi + \tau\left(\mathrm{KL}(\pi^1 \mid P) + \mathrm{KL}(\pi^2 \mid Q)\right).$$

Now, by lower semicontinuity of the KL divergence and continuity of the marginal-taking operation

$$\mathrm{KL}(\pi^1 \mid P) = \mathrm{KL}((\lim_n \pi_{k_n})^1 \mid P) = \mathrm{KL}(\lim_n \pi^1_{k_n} \mid P) \leq \lim_n \mathrm{KL}(\pi^1_{k_n} \mid P)$$

and similarly for the second marginal. Moreover, as the KL divergence is nonnegative, we get

$$F(\pi) \leq \int_{\mathcal{M}^2} d^2 \mathrm{d}\pi + \lim_n \left(\tau\left(\mathrm{KL}(\pi^1_{k_n} \mid P) + \mathrm{KL}(\pi^2_{k_n} \mid Q)\right) + \varepsilon_{k_n}\,\mathrm{KL}(\pi_{k_n} \mid P \otimes Q)\right).$$

It remains to examine the first term. Write

$$\int_{\mathcal{M}^2} \widehat{C}_{k_n}\,\mathrm{d}\pi_{k_n} = \int_{\mathcal{M}^2} d^2\,\mathrm{d}\pi_{k_n} + \int_{\mathcal{M}^2} (\widehat{C}_{k_n} - d^2)\,\mathrm{d}\pi_{k_n}.$$

By the uniform convergence of $\widehat{C}_k$ to d^2 and the uniform bound on $m(\pi_{k_n})$, as $n \to +\infty$,

$$\left| \int_{\mathcal{M}^2} (\widehat{C}_{k_n} - d^2) \, \mathrm{d}\pi_{k_n} \right| \leq \|\widehat{C}_{k_n} - d^2\|_\infty \cdot m(\pi_{k_n}) \longrightarrow 0.$$

Since d^2 is continuous on the compact $\mathcal{M}^2$ and $\pi_{k_n} \rightharpoonup \pi$, the weak convergence gives

$$\int_{\mathcal{M}^2} d^2 \, \mathrm{d}\pi_{k_n} \longrightarrow \int_{M^2} d^2 \, \mathrm{d}\pi.$$

Hence,

$$\lim_n \int_{\mathcal{M}^2} \widehat{C}_{k_n} \, \mathrm{d}\pi_{k_n} = \int_{\mathcal{M}^2} \mathrm{d}^2 \, \mathrm{d}\pi.$$

All in all, we get that

$$F(\pi) \leq \lim_n \left(\int_{\mathcal{M}^2} \widehat{C}_{k_n} \, \mathrm{d}\pi_{k_n} + \tau \left(\mathrm{KL}(\pi^1_{k_n} \mid P) + \mathrm{KL}(\pi^2_{k_n} \mid Q) \right) + \varepsilon_{k_n} \, \mathrm{KL}(\pi_{k_n} \mid P \otimes Q) \right).$$

As ε_k was chosen to realize the $\liminf$ of $F_\varepsilon(\pi_\varepsilon)$, a property inherited by its subsequence ε_{k_n}, we get

$$\mathrm{UOT}^\tau_{d^2}(P, Q) \leq F(\pi) \leq \lim_{n \to \infty} F_{k_n}(\pi_{k_n}) = \liminf_{\varepsilon \to 0} F_\varepsilon(\pi_\varepsilon) = \liminf_{\varepsilon \to 0} \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon, \varepsilon}(P, Q),$$

and the lower bound follows.

We are now able to conclude the proof, by combining the lower and upper bounds

$$\mathrm{UOT}^\tau_{d^2}(P, Q) \leq \liminf_{\varepsilon \to 0} \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon, \varepsilon}(P, Q) \leq \limsup_{\varepsilon \to 0} \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon, \varepsilon}(P, Q) \leq \mathrm{UOT}^\tau_{d^2}(P, Q).$$

Equality must hold everywhere so $\lim_{\varepsilon \to 0} \mathrm{UOT}^\tau_{\widehat{C}_\varepsilon, \varepsilon}(P, Q)$ exists and is equal to $\mathrm{UOT}^\tau_{d^2}(P, Q)$.

Case 2: $\tau = +\infty$. In the balanced case where P and Q are probability measures and we impose $\pi^1 = P$, $\pi^2 = Q$ (formally $\tau = \infty$), the marginal KL terms force the marginals to match exactly. The main adaptation from the unbalanced case is to make sure that all the necessary operations can be done in the set of couplings $\mathcal{U}(P, Q)$. This is a consequence of the following remarks :

1. As $\mathcal{M}$ is compact and d^2 is continuous, an optimal coupling exists for the balanced OT problem (Santambrogio, 2015, Thm. 1.4). Hence $\pi^\star$ is in fact a coupling.
2. For the upper bound, by the block approximation Lemma A.2, we get that the π_η are couplings, which ensures that the proof goes through.
3. For the lower bound, one observes that regularized entropic OT admits a minimizer and that it is a coupling (see, e.g., Nutz (2021, Theorem 4.2)) i.e. $(\pi_\varepsilon)_{\varepsilon > 0}$ is in fact a family of couplings. As a result, so are the sequence $(\pi_k)_{k \geq k_0}$ and its subsequence $(\pi_{k_n})_n$ as well as the limit π of that subsequence. This ensures that the proof goes through.

□

A.2 Proof of Proposition 3.2

We will in fact show that the claimed result holds for any compact, connected symmetric space. This implies as a corollary that it holds for the sphere $\mathbb{S}^2$, since the latter is a symmetric space (Helgason, 1979; Eschenburg, 1997).

Proposition A.3 (Sinkhorn divergence conditions)**.** *Let $\mathcal{M}$ be a compact, connected symmetric space. Then:*

(i) $\mathcal{M}$ is compact;

(ii) the time-ε cost function $\widehat{C}_\varepsilon(x, y) := -\varepsilon \log \mathcal{H}_{\varepsilon/4}(x, y)$ is Lipschitz continuous with respect to the geodesic distance d;

(iii) the kernel $K_\varepsilon(x,y) = e^{-\widehat{C}_\varepsilon(x,y)/\varepsilon} = \mathcal{H}_{\varepsilon/4}(x,y)$ is positive definite and universal.

Proof. (i) Compactness is assumed in the hypothesis.

(ii) The heat kernel $\mathcal{H}_\varepsilon$ on a connected manifold is smooth and strictly positive for $\varepsilon > 0$ (Grigoryan, 2009, Thm. 7.20, Cor. 8.12). Therefore $\widehat{C}_\varepsilon$ is smooth on $\mathcal{M}^2$. By compactness of $\mathcal{M}$, its Riemannian gradient is globally bounded, implying that $\widehat{C}_\varepsilon$ is Lipschitz continuous with respect to the geodesic distance (see Boumal (2023, Prop. 10.43)).

(iii) Positive definiteness and universality follow from results in Steinert et al. (2025). More precisely, therein, kernels on symmetric spaces are defined and studied through their associated *spectral densities*. There, Lemma 1 establishes that non-negativity of the spectral density implies that the associated kernel is positive definite, which holds for the heat kernel by way of an explicit formula available for the spectral density. Moreover, Theorem 3 gives conditions for which a kernel is $\mathcal{C}_c$-universal (i.e. local uniform approximation), as discussed in Micchelli et al. (2006) and required in (Feydy et al., 2019); moreover, it is explicitly stated that this applies to the heat kernel. This is precisely what we need for our purposes, concluding the proof. □

In particular, taking $\mathcal{M} = \mathbb{S}^2$, all three conditions are satisfied, so that by Séjourné et al. (2019, Theorem 5 & 6), the associated Sinkhorn divergence $\hat{S}^\tau_\varepsilon$ defines a symmetric, positive definite, smooth function that is convex in each of its input variables and metrizes the convergence in law.

A.3 Proof of Proposition 3.4

Our integration convention follows common practice in the computational spherical-harmonics literature: integrals are taken without normalization (e.g., $SO(3)$ has volume $8\pi^2$). For a comprehensive treatment, see (Driscoll & Healy, 1994; Faraut, 2008).

Proof. Fix $x \in \mathbb{S}^2$ and recall that $\vec{n} = (0,0,1)^\top$ is the north pole. Using the fact that averaging over $\psi \in [0, 2\pi)$ does not change the value, we have

$$Kf(x) = \frac{1}{2\pi}\int_0^{2\pi} Kf(x)\, d\psi = \frac{1}{2\pi}\int_0^{2\pi}\int_{\mathbb{S}^2} K(x,y) f(y)\, \sigma(dy)\, d\psi,$$

where σ denotes the surface measure on $\mathbb{S}^2$. We now parametrize the sphere via the Euler-angle representation of rotations. Let $R(\phi,\theta,\psi) \in \mathrm{SO}(3)$ denote a rotation with Euler angles (ϕ,θ,ψ); then the map

$$R(\phi,\theta,\psi) \mapsto y(\phi,\theta) = R(\phi,\theta,\psi)\vec{n} \in \mathbb{S}^2$$

covers $\mathbb{S}^2$ exactly once as ϕ and θ vary while ψ is held fixed. Under this parametrization, the surface element satisfies $\sigma(dy) = \sin\theta\, d\phi\, d\theta$. Applying Fubini's theorem, one gets

$$\int_0^{2\pi}\int_{\mathbb{S}^2} K(x,y) f(y)\, \sigma(dy)\, d\psi = \int_0^{2\pi}\int_0^{\pi}\int_0^{2\pi} K\big(x, R(\phi,\theta,\psi)\vec{n}\big) f\big(R(\phi,\theta,\psi)\vec{n}\big)\sin\theta\, d\phi\, d\theta\, d\psi.$$

The right-hand side is precisely the integral over $\mathrm{SO}(3)$ with respect to the (unnormalized) volume measure in Euler coordinates thus

$$Kf(x) = \frac{1}{2\pi}\int_{\mathrm{SO}(3)} K(x, R\vec{n}) f(R\vec{n})\, dR. \tag{7}$$

Now, using the radial symmetry of the kernel $K(x,y) = K_0(\langle x, y\rangle)$, one has

$$K(x, R\vec{n}) = K_0\big(\langle x, R\vec{n}\rangle\big) = K_0\big(\langle R^{-1}x, \vec{n}\rangle\big) = K(R^{-1}x, \vec{n}),$$

so that from (7), one gets

$$Kf(x) = \frac{1}{2\pi}\int_{\mathrm{SO}(3)} K(x, R\vec{n}) f(R\vec{n})\, dR = \frac{1}{2\pi}\int_{\mathrm{SO}(3)} f(R\vec{n})\, K(R^{-1}x, \vec{n})\, dR = \frac{1}{2\pi}(f \star \tilde{K})(x),$$

with $\tilde{K}(z) = K(z, \vec{n})$, concluding the proof. □

B SPHERICAL HARMONICS

Coordinates on $\mathbb{S}^2$. Points on the two-sphere $\mathbb{S}^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\}$ are parametrized by colatitude $\theta \in [0, \pi]$ and longitude $\varphi \in [0, 2\pi)$, with the identification

$$x = \big(\cos\varphi \sin\theta,\ \sin\varphi \sin\theta,\ \cos\theta\big)^\top \in \mathbb{R}^3.$$

The rotation group $\mathrm{SO}(3)$. The symmetry group of $\mathbb{S}^2$ is the *special orthogonal group*

$$\mathrm{SO}(3) = \big\{R \in \mathbb{R}^{3\times 3} : R^\top R = I,\ \det R = 1\big\},$$

consisting of all proper (orientation-preserving) rotations of $\mathbb{R}^3$. Any $R \in \mathrm{SO}(3)$ can be factored using the *Euler angles* $\varphi \in [0, 2\pi)$, $\theta \in [0, \pi]$, $\psi \in [0, 2\pi)$ as

$$R(\phi, \theta, \psi) = R_z(\phi)\, R_y(\theta)\, R_z(\psi),$$

where R_z and R_y denote elementary rotations around the z- and y-axes:

$$R_z(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad R_y(\alpha) = \begin{pmatrix} \cos\alpha & 0 & \sin\alpha \\ 0 & 1 & 0 \\ -\sin\alpha & 0 & \cos\alpha \end{pmatrix}.$$

Unlike translations in $\mathbb{R}^n$, rotations do not commute in general, making $\mathrm{SO}(3)$ non-abelian.

L^2 inner product. The natural inner product on $\mathbb{S}^2$ is

$$\langle u, v\rangle_{L^2(\mathbb{S}^2)} = \int_{\mathbb{S}^2} u\, \bar{v}\, \mathrm{d}\Omega = \int_0^{2\pi}\!\!\int_0^{\pi} u(\theta,\varphi)\, \overline{v(\theta,\varphi)}\, \sin\theta\, \mathrm{d}\theta\, \mathrm{d}\varphi,$$

where $\mathrm{d}\Omega = \sin\theta\, \mathrm{d}\theta\, \mathrm{d}\varphi$ is the rotation-invariant surface measure on $\mathbb{S}^2$. This defines the Hilbert space $L^2(\mathbb{S}^2)$ of square-integrable functions on the sphere.

Spherical harmonics. The *spherical harmonics* are defined for degree $\ell \geq 0$ and order $|m| \leq \ell$ by

$$Y_{\ell m}(\theta, \varphi) = N_{\ell m}\, P_\ell^m(\cos\theta)\, e^{im\varphi}, \qquad N_{\ell m} = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}},$$

where P_ℓ^m are the associated Legendre polynomials. The normalization $N_{\ell m}$ is chosen so that the family $\{Y_{\ell m}\}_{\ell\geq 0,\, |m|\leq \ell}$ is orthonormal:

$$\langle Y_{\ell m},\, Y_{\ell' m'}\rangle_{L^2(\mathbb{S}^2)} = \delta_{\ell\ell'}\, \delta_{mm'}.$$

Among all possible orthonormal bases of $L^2(\mathbb{S}^2)$, the spherical harmonics are distinguished by their equivariance under $\mathrm{SO}(3)$: applying any rotation $R \in \mathrm{SO}(3)$ to $Y_{\ell m}$ produces a linear combination of harmonics of the *same* degree ℓ,

$$Y_{\ell m}(R^{-1}\,\cdot) = \sum_{|m'|\leq \ell} D^\ell_{m'm}(R)\, Y_{\ell m'},$$

where $D^\ell(R)$ is the Wigner D-matrix of degree ℓ. The family $\{Y_{\ell m}\}$ is moreover complete, so any $f \in L^2(\mathbb{S}^2)$ admits the *spherical harmonic expansion*

$$f(\theta,\varphi) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell} \widehat{f}_{\ell m}\, Y_{\ell m}(\theta,\varphi), \qquad \widehat{f}_{\ell m} = \mathcal{F}[f](\ell, m) \coloneqq \langle f,\, Y_{\ell m}\rangle_{L^2(\mathbb{S}^2)}.$$

Spherical convolution. On $\mathbb{S}^2$, convolution is defined by averaging over rotations rather than translations. Following Driscoll & Healy (1994), this leads to the spherical convolution operator

$$(f \star \kappa)(x) = \int_{\mathrm{SO}(3)} f(R\vec{n})\, \kappa(R^{-1}x)\, \mathrm{d}R,$$

which admits a convolution theorem in the spherical harmonic domain.

Proposition 3.4 shows that our radial kernel operator coincides exactly with spherical convolution (up to normalization), enabling fast Sinkhorn iterations via fast SHTs. Other uses of spherical convolution for neural networks include Kondor & Trivedi (2018) which studies group convolutions on homogeneous spaces including $\mathbb{S}^2$, while Sosnovik et al. (2021) develop scale-equivariant layers via hierarchical sampling.

C Implementation Details

Our implementation uses the differentiable spherical harmonic transform libraries `torch-harmonics` (Bonev et al., 2023) and `s2fft` (Price & McEwen, 2024), built on top of `PyTorch` (Paszke et al., 2019) and `JAX` (Bradbury et al., 2018).

Algorithm 1 Spherical Harmonic Sinkhorn

Require: $\mathbf{p}, \mathbf{q} \in \mathbb{R}_+^n$ (source and target marginals), $\varepsilon > 0$ (regularization), $\tau \in (0, +\infty]$ (unbalanced parameter), $T \in \mathbb{N}$ (max iterations), $L \in \mathbb{N}$ (band-limit), $\{a_i\}$ (area weights of the discretization), $\delta > 0$ (convergence threshold), $C \in \mathbb{N}$ (check frequency)

$t \leftarrow \varepsilon/4$
$h_\ell \leftarrow e^{-t\ell(\ell+1)}, \quad \ell = 0, \dots, L-1$
$\mathbf{u} \leftarrow \mathbf{1}$, $\mathbf{v} \leftarrow \mathbf{1}$
for $k = 1, \dots, T$ **do**
  $\mathbf{u}_{\text{prev}} \leftarrow \mathbf{u}$, $\mathbf{v}_{\text{prev}} \leftarrow \mathbf{v}$
  if $\tau = +\infty$ **then**
    $\mathbf{u} \leftarrow \mathbf{p} \oslash \mathrm{HeatConv}(\mathbf{v}; \{h_\ell\})$ {with $\oslash$ being the element-wise division}
    $\mathbf{v} \leftarrow \mathbf{q} \oslash \mathrm{HeatConv}(\mathbf{u}; \{h_\ell\})$
    $s \leftarrow \sum_i u_i$
    $\mathbf{u} \leftarrow \mathbf{u}/s, \quad \mathbf{v} \leftarrow s\,\mathbf{v}$
  else
    $\mathbf{u} \leftarrow \big(\mathbf{p} \oslash \mathrm{HeatConv}(\mathbf{v}; \{h_\ell\})\big)^{\tau/(\tau+\varepsilon)}$
    $\mathbf{v} \leftarrow \big(\mathbf{q} \oslash \mathrm{HeatConv}(\mathbf{u}; \{h_\ell\})\big)^{\tau/(\tau+\varepsilon)}$
    $c \leftarrow \frac{\tau}{2}\left(\log\left(\sum_i a_i\, p_i\, u_i^{-\varepsilon/\tau}\right) - \log\left(\sum_i a_i\, q_i\, v_i^{-\varepsilon/\tau}\right)\right)$
    $\mathbf{u} \leftarrow e^{c/\varepsilon}\,\mathbf{u}, \quad \mathbf{v} \leftarrow e^{-c/\varepsilon}\,\mathbf{v}$
  end if
  if $k \bmod C = 0$ **then**
    $e_{\text{cvg}} \leftarrow \texttt{ConvergenceCriterion}(\mathbf{u}, \mathbf{v}, \mathbf{u}_{\text{prev}}, \mathbf{v}_{\text{prev}}, \mathbf{p})$ {See Appendix C.3}
    if $e_{\text{cvg}} < \delta$ **then**
      break
    end if
  end if
end for
$\phi \leftarrow \varepsilon \log \mathbf{u}, \quad \psi \leftarrow \varepsilon \log \mathbf{v}$
if $\tau = +\infty$ **then**
  return $\sum_i a_i \big[\phi_i\, p_i + \psi_i\, q_i\big]$
else
  return $\sum_i a_i \big[\tau(1 - e^{-\phi_i/\tau})\, p_i + \tau(1 - e^{-\psi_i/\tau})\, q_i\big]$
end if

C.1 Discretizations of $\mathbb{S}^2$

In practice, functions on $\mathbb{S}^2$ are represented on a finite grid of sampling points (θ_i, ϕ_j). We briefly review several such discretization schemes, illustrated in Figures 5 and 6. Certain schemes admit an exact sampling theorem: for bandlimited signals, i.e., functions whose expansion in the spherical harmonics is supported on degrees $\ell \leq L$, a sufficiently dense sampling guarantees that both the forward and inverse Spherical Harmonic Transforms are exact.

- **Equiangular (Driscoll-Healy (Driscoll & Healy, 1994) / McEwen-Wiaux (McEwen & Wiaux, 2011) / Clenshaw-Curtis (Clenshaw & Curtis, 1960)).** Samples are placed on a regular grid equally spaced in longitude and at equiangular latitudes. The resolution is parametrized by a band-limit L. These grids admit exact sampling theorems, which require $\mathcal{O}(L^2)$ samples.
- **Gauss-Legendre.** Latitudinal samples are placed at the L roots of the Legendre polynomial P_L, yielding an exact sampling theorem. As with equiangular grids, the resolution is parametrized by L and the total number of samples is $\mathcal{O}(L^2)$.

- **Lobatto (Clenshaw-Curtis with poles).** A variant of equiangular sampling that includes the poles as quadrature nodes, using the zeros of $(1 - x^2)P'_{L-1}(x)$ in the latitudinal direction. It shares the same band-limit parametrization and sampling requirements as the schemes above.

- **HEALPix (Górski et al., 2005).** The sphere is partitioned into $N = 12N_{\text{side}}^2$ equal-area pixels arranged in iso-latitude rings. The resolution is controlled by N_{side}. No exact sampling theorem is available, and N is an independent resolution parameter.

For exact iso-latitudinal schemes (Clenshaw-Curtis, Gauss-Legendre, Lobatto), spherical harmonic transforms exploit a separation-of-variables structure. A direct implementation scales as $\mathcal{O}(L^4)$, while the standard semi-naïve algorithm reduces this to $\mathcal{O}(L^3)$ by combining FFTs in longitude with projections onto associated Legendre polynomials. Faster algorithms with complexity $\mathcal{O}(L^2 \log^2 L)$, based on fast polynomial transforms, have been proposed (Healy Jr et al., 2003) and in some cases extend beyond specific sampling schemes (Kunis & Potts, 2003). However, these approaches typically require additional stabilization procedures whose numerical robustness and complexity are not fully understood in general. As a result, despite their favorable asymptotic scaling, they are not yet as reliable in practice. Consequently, most implementations (e.g., Price & McEwen (2024)) rely on the semi-naïve $\mathcal{O}(L^3)$ algorithms for both forward and inverse transforms, which offer a more robust trade-off between efficiency and numerical stability. Similar considerations apply to HEALPix, where fast methods exist (Kunis & Potts, 2003; Drake & Wright, 2020) but practical computations often favor stable, moderately super-quadratic algorithms.

In our experiments, we use the Clenshaw-Curtis grid for an equiangular sampling and the HEALPix grid for an equal-area sampling scheme as it is widely used in the climate science community. However our implementation is flexible and can be easily adapted to other discretization schemes.

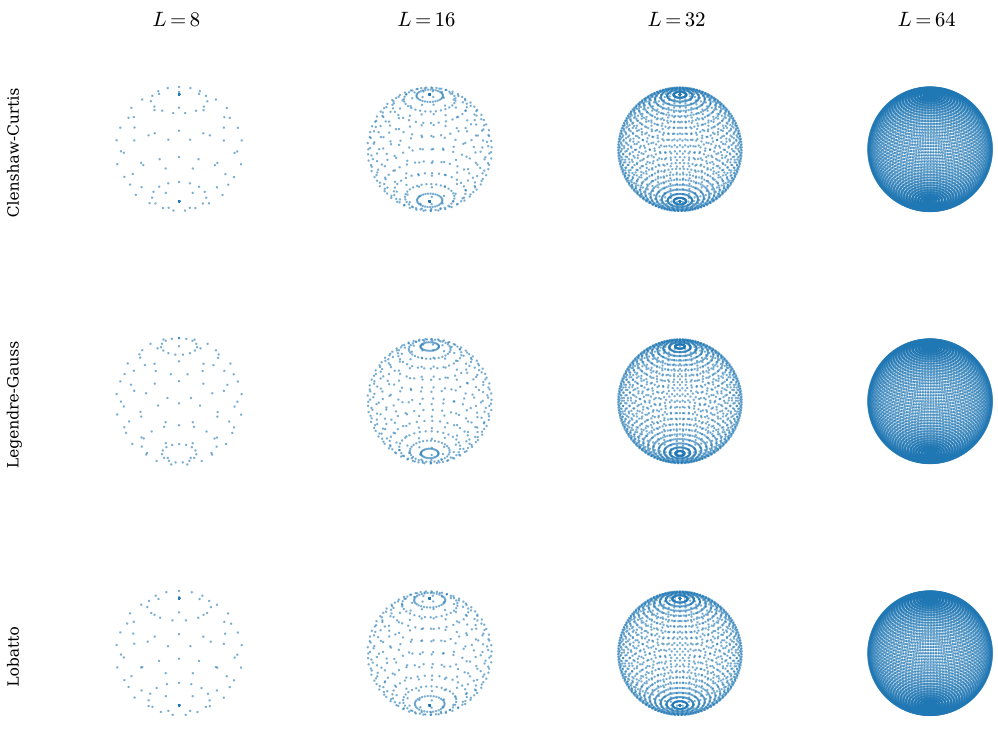


Figure 5: Clenshaw-Curtis, Gauss-Legendre and Lobatto discretization schemes for varying L.

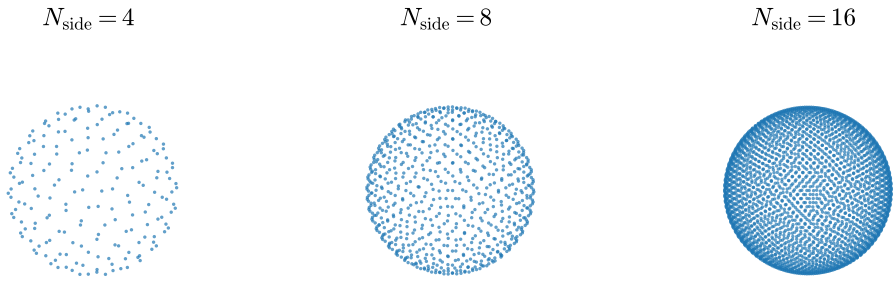


Figure 6: HEALPix discretization scheme for varying N_{side}.

C.2 Discrete OT on the Sphere

Let $\{x_i\}_{i=1}^n$ be a discretization of $\mathbb{S}^2$ with area weights $\mathbf{a} \in \mathbb{R}_+^n$, $\mathbf{a}^\top \mathbf{1} = 4\pi$, so that $\int_{\mathbb{S}^2} p \, \mathrm{d}\sigma \approx \mathbf{a}^\top \mathbf{p}$ where $\mathbf{p} = (p(x_1), \dots, p(x_n))$. A transport plan between two positive measures $\mathbf{p}, \mathbf{q} \in \mathbb{R}_+^n$ is represented as a matrix $\boldsymbol{\pi} \in \mathbb{R}_+^{n \times n}$, where π_{ij} denotes the density of mass transported from x_i to x_j. This gives the area-weighted discrete entropic OT problem $\mathrm{OT}_{C,\varepsilon}^{\tau}(\mathbf{p}, \mathbf{q})$ as

$$\min_{\boldsymbol{\pi} \geq 0} \sum_{i,j} C(x_i, x_j)\, \pi_{ij}\, a_i a_j + \varepsilon\, \mathrm{KL}_{\mathbf{a}}(\boldsymbol{\pi} \mid \mathbf{p} \otimes \mathbf{q}) + \tau \left(\mathrm{KL}_{\mathbf{a}}(\boldsymbol{\pi}\mathbf{a} \mid \mathbf{p}) + \mathrm{KL}_{\mathbf{a}}(\boldsymbol{\pi}^\top \mathbf{a} \mid \mathbf{q}) \right), \tag{8}$$

where

$$\mathrm{KL}_{\mathbf{a}}(\boldsymbol{\pi} \mid \mathbf{p} \otimes \mathbf{q}) = \sum_{i,j} \pi_{ij} \left(\log \frac{\pi_{ij}}{p_i q_j} - 1 \right) a_i a_j$$

and

$$\mathrm{KL}_{\mathbf{a}}(\boldsymbol{\pi}\mathbf{a} \mid \mathbf{p}) = \sum_i (\boldsymbol{\pi}\mathbf{a})_i \left(\log \frac{(\boldsymbol{\pi}\mathbf{a})_i}{p_i} - 1 \right) a_i.$$

We now show that (8) is equivalent to the standard discrete entropic OT problem on the probability mass vectors $\bar{\mathbf{p}} = \mathrm{diag}(\mathbf{a})\mathbf{p}$, and $\bar{\mathbf{q}} = \mathrm{diag}(\mathbf{a})\mathbf{q}$. Let $\boldsymbol{\gamma} = \mathrm{diag}(\mathbf{a})\, \boldsymbol{\pi}\, \mathrm{diag}(\mathbf{a})$, then (8) is equivalent to

$$\min_{\boldsymbol{\gamma} \geq 0} \sum_{ij} C(x_i, x_j)\, \gamma_{ij} + \varepsilon\, \mathrm{KL}_{\mathbf{1}}(\boldsymbol{\gamma} \mid \bar{\mathbf{p}} \otimes \bar{\mathbf{q}}) + \tau \big(\mathrm{KL}_{\mathbf{1}}(\boldsymbol{\gamma}\mathbf{1} \mid \bar{\mathbf{p}}) + \mathrm{KL}_{\mathbf{1}}(\boldsymbol{\gamma}^\top \mathbf{1} \mid \bar{\mathbf{q}}) \big). \tag{9}$$

Under the change of variables $\gamma_{ij} = a_i a_j \pi_{ij}$ (i.e. $\pi_{ij} = \gamma_{ij}/(a_i a_j)$), the three terms transform as follows. The cost satisfies $\sum_{ij} C(x_i, x_j)\, \pi_{ij}\, a_i a_j = \sum_{ij} C(x_i, x_j)\, \gamma_{ij}$ directly. The regularization becomes

$$\mathrm{KL}_{\mathbf{a}}(\boldsymbol{\pi} \mid \mathbf{p} \otimes \mathbf{q}) = \sum_{ij} a_i a_j\, \pi_{ij} \left(\log \frac{\pi_{ij}}{p_i q_j} - 1 \right) = \sum_{ij} \gamma_{ij} \left(\log \frac{\gamma_{ij}}{\bar{p}_i \bar{q}_j} - 1 \right) = \mathrm{KL}_{\mathbf{1}}(\boldsymbol{\gamma} \mid \bar{\mathbf{p}} \otimes \bar{\mathbf{q}}).$$

Finally, since $(\boldsymbol{\pi}\mathbf{a})_i = (\boldsymbol{\gamma}\mathbf{1})_i / a_i$,

$$\mathrm{KL}_{\mathbf{a}}(\boldsymbol{\pi}\mathbf{a} \mid \mathbf{p}) = \sum_i a_i\, (\boldsymbol{\pi}\mathbf{a})_i \left(\log \frac{(\boldsymbol{\pi}\mathbf{a})_i}{p_i} - 1 \right) = \sum_i (\boldsymbol{\gamma}\mathbf{1})_i \left(\log \frac{(\boldsymbol{\gamma}\mathbf{1})_i}{\bar{p}_i} - 1 \right) = \mathrm{KL}_{\mathbf{1}}(\boldsymbol{\gamma}\mathbf{1} \mid \bar{\mathbf{p}}),$$

and symmetrically $\mathrm{KL}_{\mathbf{a}}(\boldsymbol{\pi}^\top \mathbf{a} \mid \mathbf{q}) = \mathrm{KL}_{\mathbf{1}}(\boldsymbol{\gamma}^\top \mathbf{1} \mid \bar{\mathbf{q}})$. Hence (8) and (9) coincide.

We use Sinkhorn algorithm to solve the area-weighted discrete entropic OT problem (8). Once it has converged, we recover the optimal Kantorovich potentials $\phi = \varepsilon \log \mathbf{u}$ and $\psi = \varepsilon \log \mathbf{v}$ and compute the transport cost as

$$\mathrm{OT}_{C,\varepsilon}^{\tau}(\mathbf{p}, \mathbf{q}) = \sum_i a_i\, \tau \Big(1 - e^{-\phi_i/\tau} \Big)\, p_i + \sum_i a_i\, \tau \Big(1 - e^{-\psi_i/\tau} \Big)\, q_i. \tag{10}$$

This follows from strong duality for unbalanced entropic OT (Chizat et al., 2018). In the balanced case $\tau = \infty$, the formula reduces to the standard dual $\sum_i a_i\, \phi_i\, p_i + \sum_i a_i\, \psi_i\, q_i$.

C.3 Convergence Criterion

In the balanced case, we monitor convergence via the relative left marginal violation. After each pair of Sinkhorn updates, we evaluate

$$e_{\text{cvg}} = \frac{\left\|\mathbf{u} \odot \text{HeatConv}(\mathbf{v};\, \{h_\ell\}) - \mathbf{p}\right\|_2}{\|\mathbf{p}\|_2}, \tag{11}$$

and declare convergence when $e_{\text{cvg}} < \delta$. To reduce overhead, this check is performed every C iterations. We use $C = 10$ throughout our experiments.

In the unbalanced setting, since the marginal constraints are relaxed and need not be satisfied, we monitor instead the relative L^2 distance between two iterations. More precisely, we define

$$e_{\text{cvg}} = \frac{1}{2}\left(\frac{\|\log(u^{(k+1)}) - \log(u^{(k)})\|_2}{\|\log(u^{(k)})\|_2} + \frac{\|\log(v^{(k+1)}) - \log(v^{(k)})\|_2}{\|\log(v^{(k)})\|_2}\right). \tag{12}$$

C.4 Numerical instabilities

In `PyTorch` with `float64` arithmetic, the smallest representable positive number is approximately 4.94×10^{-324}. On $\mathbb{S}^2$, the squared-distance matrix $C_{ij} = d(x_i, x_j)^2$ has entries as large as π^2. The Sinkhorn algorithm requires evaluating the Gibbs kernel $\mathcal{K}_{ij} = e^{-C_{ij}/\varepsilon}$, whose entries underflow to zero as soon as $d(x_i, x_j)^2/\varepsilon \geq \ln\big(1/(4.94 \times 10^{-324})\big) \approx 744.4$. Antipodal pairs are the first to be affected, at $\varepsilon \leq \pi^2/744.4 \approx 0.013$. As ε decreases further, the effective support of the kernel around each point shrinks: progressively closer pairs underflow in turn until $\mathcal{K}$ is effectively diagonal. At that point Sinkhorn iterations can no longer redistribute mass and the algorithm degenerates.

A distinct instability arises from the truncated spherical harmonic representation. The Sinkhorn in Heat convolution step evaluates $\mathcal{F}^{-1}(\mathcal{F}[u](l,m) \cdot e^{-\frac{\varepsilon}{4}\ell(\ell+1)})$, which approximates the convolution by the heat kernel by a finite-degree expansion up to L. Although the exact heat kernel is a strictly positive operator, its degree-L truncation can take small negative values when $\varepsilon^{-1/2} \gtrsim L$, i.e. when the kernel is too narrow to be resolved on the SH grid. Even a single negative entry in the reconstructed convolution suffices to make the Sinkhorn update undefined. Note that clamping the output values to a small positive scalar introduces too big errors and does not prevent the algorithm from failing.

D Additional Numerical Experiments

D.1 Runtime comparison

We use the same setup as in the runtime comparison of the main paper (random probability measures on $\mathbb{S}^2$, single NVIDIA A40 GPU, `float64` precision, 5 runs per configuration), but vary the regularization parameter ε across multiple values instead of fixing it to 0.1. Figure 7 shows that the speed and scalability and advantage of SHOT observed in Figure 1 is consistent across values of ε and grid types.

Unbalanced case. We also repeat the benchmark in the unbalanced setting, keeping the setup of the main paper. The second measure is rescaled to 0.8 times the mass of the first one, and the marginal relaxation is set to $\tau = 10$. We compare SHOT (ours) and Sinkhorn (with explicit kernel matrix convolutions), both run in the log domain for numerical stability; the other baselines of the main paper (SSW, LSSOT, Cholesky and the exact linear program) are not applicable in the unbalanced setting.

Figure 8 shows the same qualitative picture as in the balanced case: Sinkhorn, which builds an explicit $N \times N$ kernel matrix, runs out of memory beyond $L = 128$ and $N_{\text{side}} = 32$, whereas SHOT scales to the largest resolutions considered. On the Clenshaw-Curtis grid, the runtime of SHOT remains essentially flat in L, while Sinkhorn degrades sharply as soon as the kernel matrix becomes large, and is already slower at $L = 128$; at low resolution, where the dense distance matrix still fits in memory, it remains faster. On the HEALPix grid, we again observe the overhead of the `s2fft` transforms (discussed in Section 4.1), which makes SHOT slower than Sinkhorn in the range of resolutions where the latter does not run out of memory.

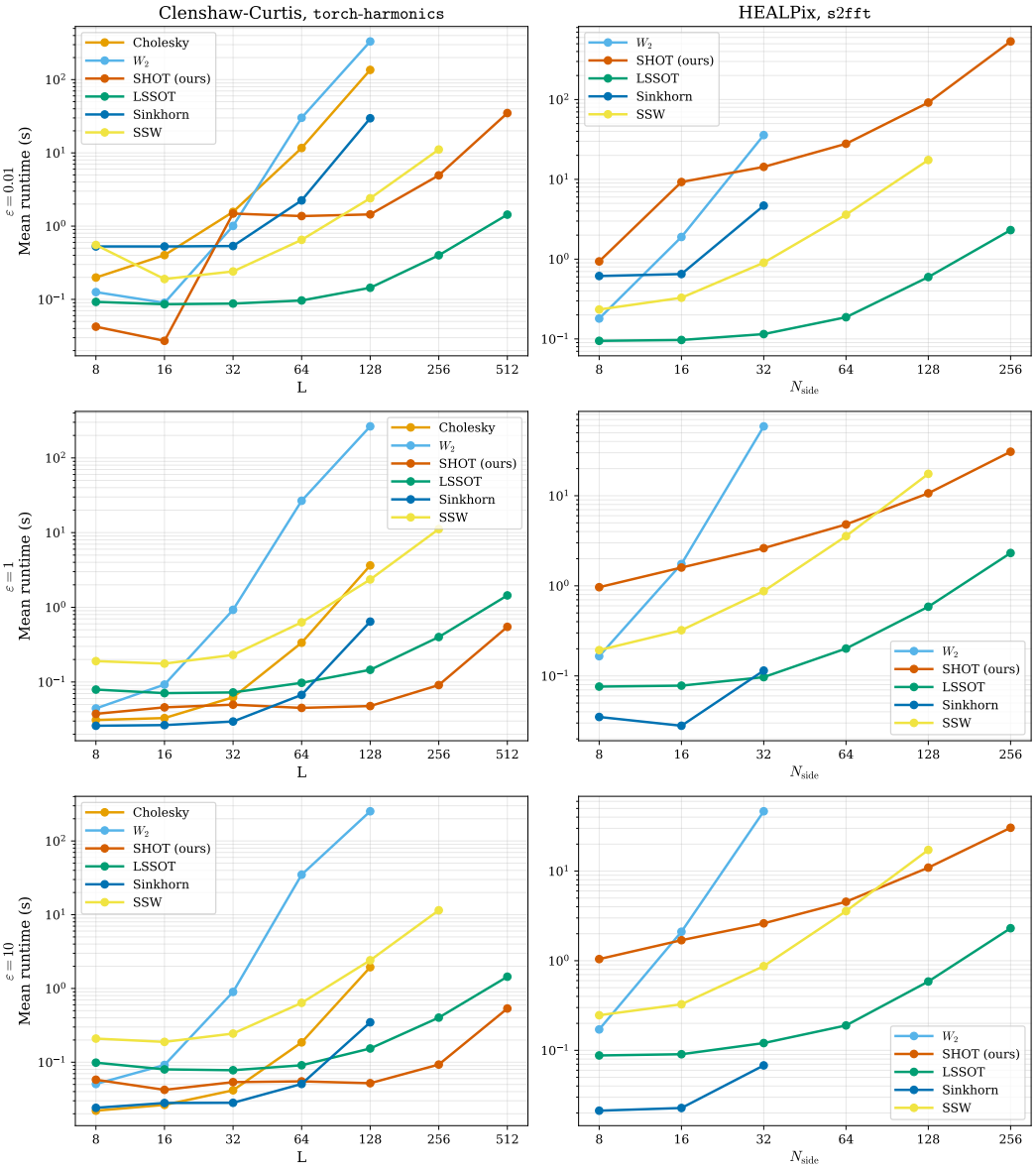


Figure 7: Runtime comparison across methods, grid types, and regularization parameters ε.

D.2 Convolution on $\mathbb{S}^2$

We validate that the heat-kernel convolution and the explicit matrix convolution converge to the same result as the grid resolution increases. The input distribution is a Von Mises-Fisher (vMF) measure with concentration $\kappa = 10$ and a fixed random mean direction. All experiments are run on CPU with `float64` precision.

Kernels and normalization. The two convolutions operate with kernels at different scales and must be brought to a common scale before comparison. The *matrix* convolution applies the unnormalized geodesic kernel $K_\varepsilon(x, y) = e^{-d(x,y)^2/\varepsilon}$, which integrates to

$$C(\varepsilon) = \int_{\mathbb{S}^2} e^{-d(x,y)^2/\varepsilon} \, \mathrm{d}\sigma(y),$$

a quantity that depends on ε and goes to 0 as $\varepsilon \to 0$. The *heat* convolution applies the heat kernel H_t at time $t = \varepsilon/4$, which is a *stochastic* (probability-preserving) kernel satisfying $\int_{\mathbb{S}^2} H_t(x, y) \, \mathrm{d}\sigma(y) = 1$.

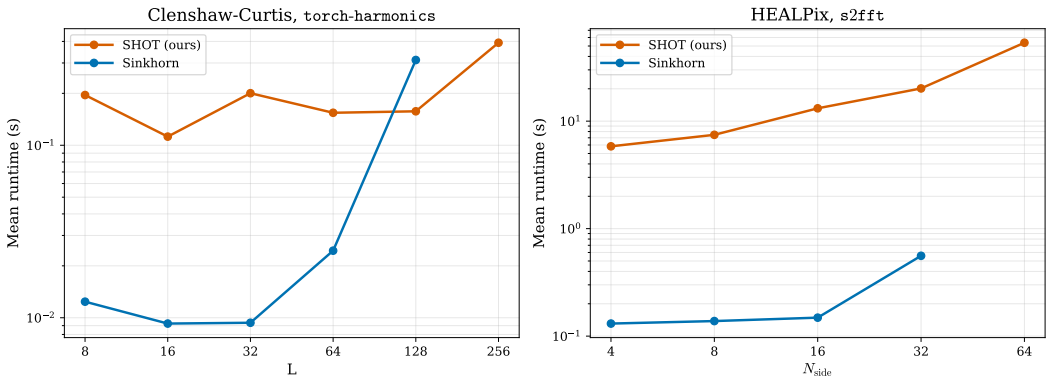


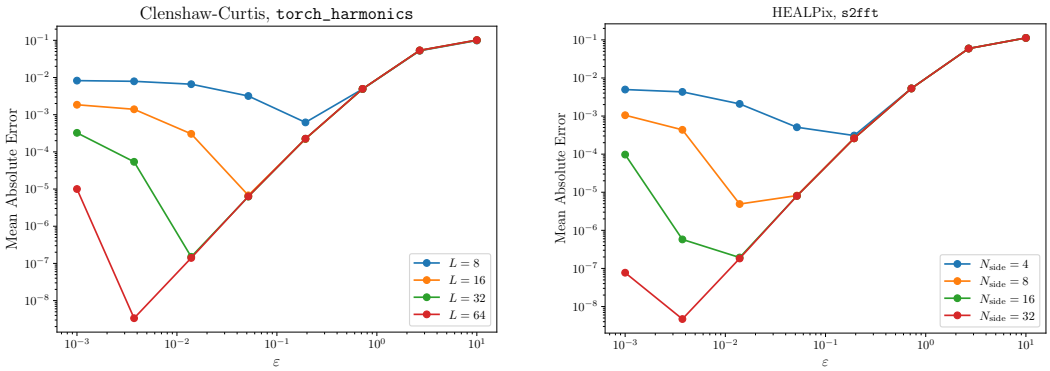
Figure 8: Unbalanced runtime comparison ($\varepsilon = 0.1$, $\tau = 10$, mass ratio 0.8): mean runtime (seconds) vs. grid resolution averaged over 5 runs; missing values indicate out-of-memory. *Left*: Clenshaw-Curtis grid. *Right*: HEALPix grid.

By Varadhan's formula, $H_{\varepsilon/4}(x, y) \approx e^{-d(x,y)^2/\varepsilon}/C(\varepsilon)$ for small ε, so the two outputs are related by

$$(H_{\varepsilon/4} * f)(x) \approx \frac{(K_\varepsilon * f)(x)}{C(\varepsilon)}.$$

Comparing them directly would therefore show a systematic amplitude difference of $C(\varepsilon)$, which grows large as $\varepsilon \to 0$, rather than a discretization error. To isolate the approximation error, we rescale the heat output by $C(\varepsilon)$ before computing the mean absolute error. Note that this rescaling is irrelevant in the Sinkhorn iterations, where scaling K by any constant is absorbed by the dual variables and leaves the optimal transport plan unchanged.

Setup and results. We consider two discretizations of $\mathbb{S}^2$: the Clenshaw-Curtis equiangular grid (via `torch-harmonics`) with band limits $L \in \{8, 16, 32, 64\}$, and the HEALPix grid (via `s2fft`) with $N_{\text{side}} \in \{4, 8, 16, 32\}$. For each $\varepsilon \in [10^{-3}, 10]$ (8 values log-spaced), we apply both convolutions and measure their mean absolute error after rescaling. As expected, the error decreases as L (resp. N_{side}) grows, and is larger at large ε where the heat-kernel approximation is less accurate. Figure 9 shows the quantitative error curves for both grids, and Figures 10 and 11 show the qualitative Mollweide maps for the Clenshaw-Curtis and HEALPix grids respectively.



(a) Clenshaw-Curtis (`torch-harmonics`), $L \in \{8, 16, 32, 64\}$.

(b) HEALPix (`s2fft`), $N_{\text{side}} \in \{4, 8, 16, 32\}$.

Figure 9: Mean absolute error between heat and matrix convolutions as a function of ε, for both grid types.

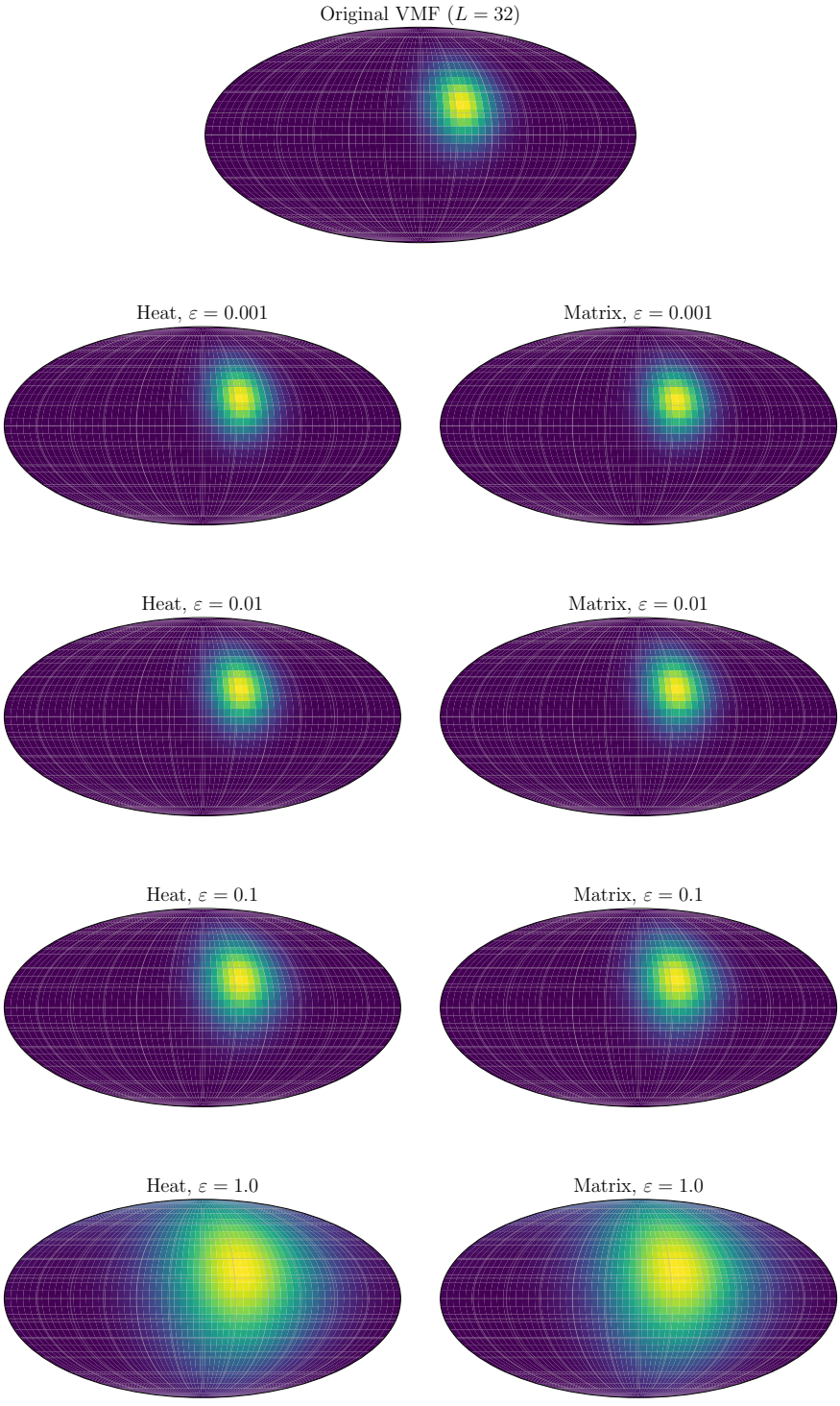


Figure 10: Mollweide projections of the heat (left column) and matrix (right column) convolutions of a vMF measure on the Clenshaw-Curtis grid ($L = 32$), for $\varepsilon \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$ (top to bottom).

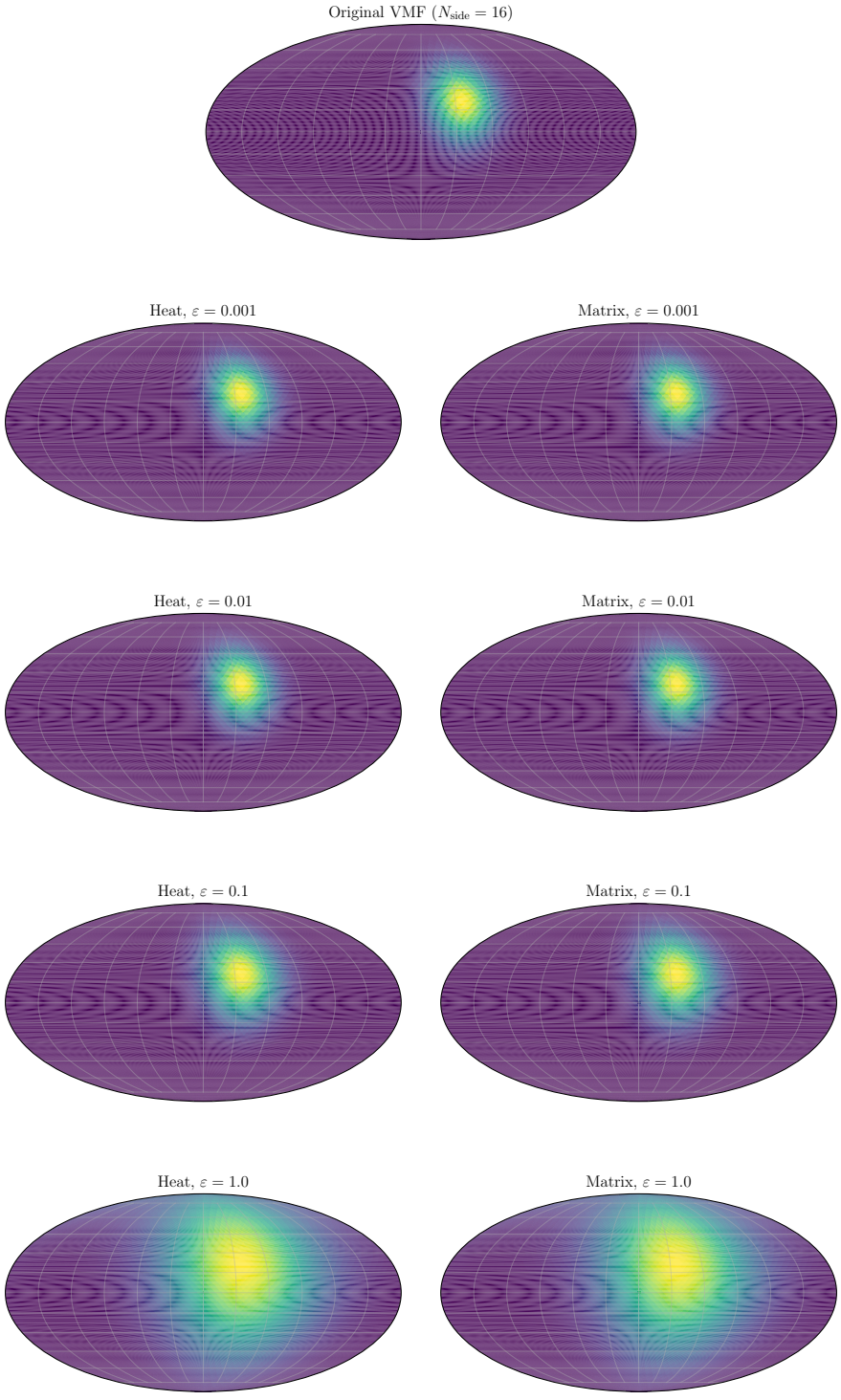


Figure 11: Mollweide projections of the heat (left column) and matrix (right column) convolutions of a vMF measure on the HEALPix grid ($N_{\text{side}} = 16$), for $\varepsilon \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$ (top to bottom).

E Details on Climate Models Comparisons

Precipitation fields are represented as discrete measures on the HEALPix grid (Górski et al., 2005) at resolution $N_{\text{side}} = 64$ ($\sim$49 152 equal-area pixels, $\approx 1^\circ$ effective resolution), which provides an equal-area pixelation of $\mathbb{S}^2$ directly compatible with our algorithm. ERA5 data is accessed via the HERA5 dataset[3], already provided in HEALPix format, while CMIP6 model outputs are retrieved from the Pangeo Google Cloud archive[4] and regridded to the same HEALPix grid via bilinear interpolation. All fields are converted to mm/day. We focus on total precipitation and analyze two climatological seasons: December-January-February (DJF) and June-July-August (JJA), each computed as a temporal average over the 2000-2014 historical period. Visual illustrations of the resulting precipitation distributions for ERA5 and one model (ACCESS-CM2) are shown in Figure 12. The choice of precipitation as the variable of interest is motivated by its critical role in the climate system and its known challenges for climate models, which often struggle to accurately represent its spatial distribution and intensity.

Gradient-based regional attribution. We partition $\mathbb{S}^2$ into five geographic regions: Tropics ($|\text{lat}| < 23.5^\circ$), ITCZ ($|\text{lat}| < 10^\circ$), NH Midlatitudes (30°–60°N), SH Midlatitudes (30°–60°S), and Arctic ($> 66.5^\circ$N), and compute the area-weighted RMS of g_m over each region as a regional bias index. These indices are reported for all eight models in both seasons in Figure 13, normalized by the global maximum across all models, regions, and seasons to preserve inter-model differences.

Results. Four physically interpretable patterns emerge from Figure 13.

Tropical and ITCZ biases are structurally season-independent. These two regions show the largest absolute gradient norms in both seasons and the smallest DJF$-$JJA differences ($|\Delta| \leq 0.02$ for all models), confirming that tropical convection deficiencies in CMIP6 are a structural property of model parameterizations rather than a seasonally forced error. This has a direct practical implication: improving tropical precipitation requires architectural changes to convection schemes, not seasonal retuning.

SH midlatitudes are uniformly well captured. The SH midlatitude column shows the lowest gradient norms across both seasons ($|\Delta| \leq 0.01$) and negligible inter-model spread, indicating that Southern Ocean precipitation is similarly represented by all models and that the unbalanced Sinkhorn divergence adds no discriminating power in this region. This is itself a useful negative result.

NH midlatitudes show a mild but consistent DJF elevation. Most models exhibit slightly larger NH midlatitude bias in DJF than JJA ($\Delta = +0.02$ for ACCESS-CM2, GFDL-ESM4, IPSL-CM6A-LR, MRI-ESM2-0; $\Delta = 0.00$ for MPI-ESM1-2-HR and CanESM5), consistent with the activation of boreal winter storm tracks and polar vortex dynamics. The signal is secondary and not universal, suggesting model-dependent sensitivity to wintertime circulation.

Sensitivity analysis of parameter τ. We ran the scalar ranking analysis (Section 4.2.2) across $\tau \in \{10, 30, 100, 300, 1000\}$, a two-order-of-magnitude sweep centered on our chosen value. The four lowest-ranked models (CanESM5, NorESM2-MM, UKESM1-0-LL, ACCESS-CM2) retain their exact rank ordering across the entire range, and the top-ranked model (GFDL-ESM4) is stable for all $\tau \geq 30$ (see Figure 14). Reshuffling is confined to a tightly-clustered top-4 group (GFDL-ESM4, MRI-ESM2-0, IPSL-CM6A-LR, MPI-ESM1-2-HR), which exchange ranks among themselves but never cross into the bottom half. Quantitatively, Kendall's τ between the $\tau = 100$ ranking and the rankings at $\tau = 10$ and $\tau = 1000$ is

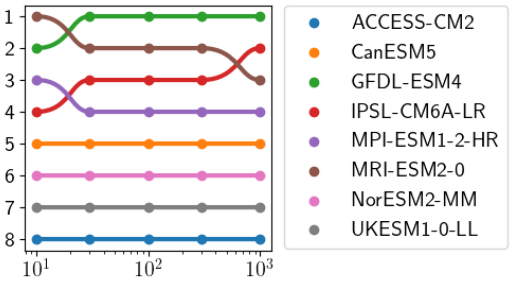


Figure 14: Sensitivity of model rankings to the parameter τ.

[3] https://orcestra-campaign.org/hera5.html
[4] https://pangeo-data.github.io/pangeo-cmip6-cloud/

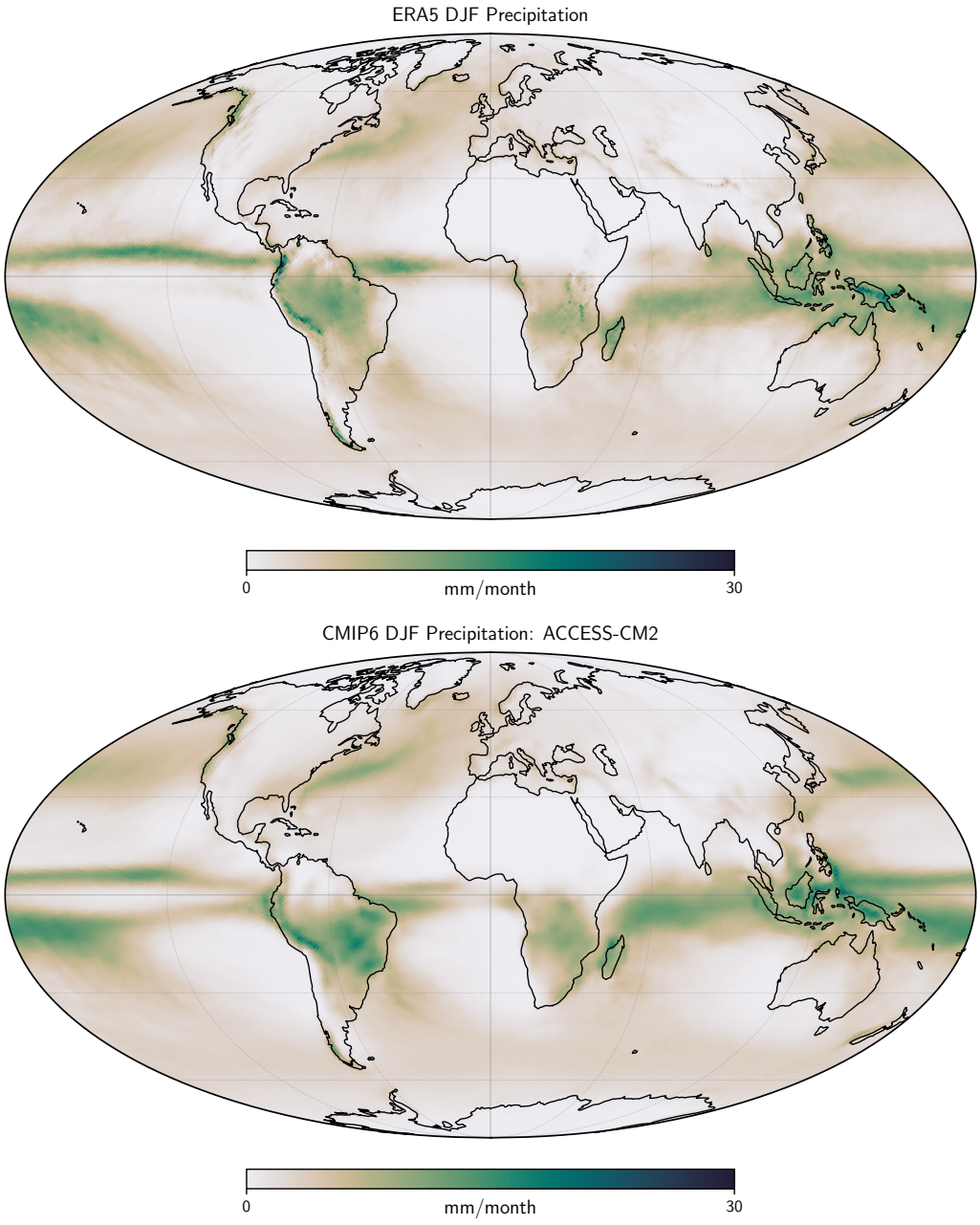


Figure 12: Illustrations of the precipitation distributions for the ERA5 reanalysis (top) and the ACCESS-CM2 model (bottom) during the DJF season.

0.86; between $\tau = 100$ and $\tau = 30$ or $\tau = 300$ it is 1.00 (identical ranking). We take this as evidence that our qualitative conclusions, which models are most/least biased relative to ERA5, are not an artifact of the specific $\tau = 100$ choice, and that the observed sensitivity is concentrated among models whose absolute scores are already close, which is the expected behavior of any ranking under near-ties rather than a sign of instability in the method.

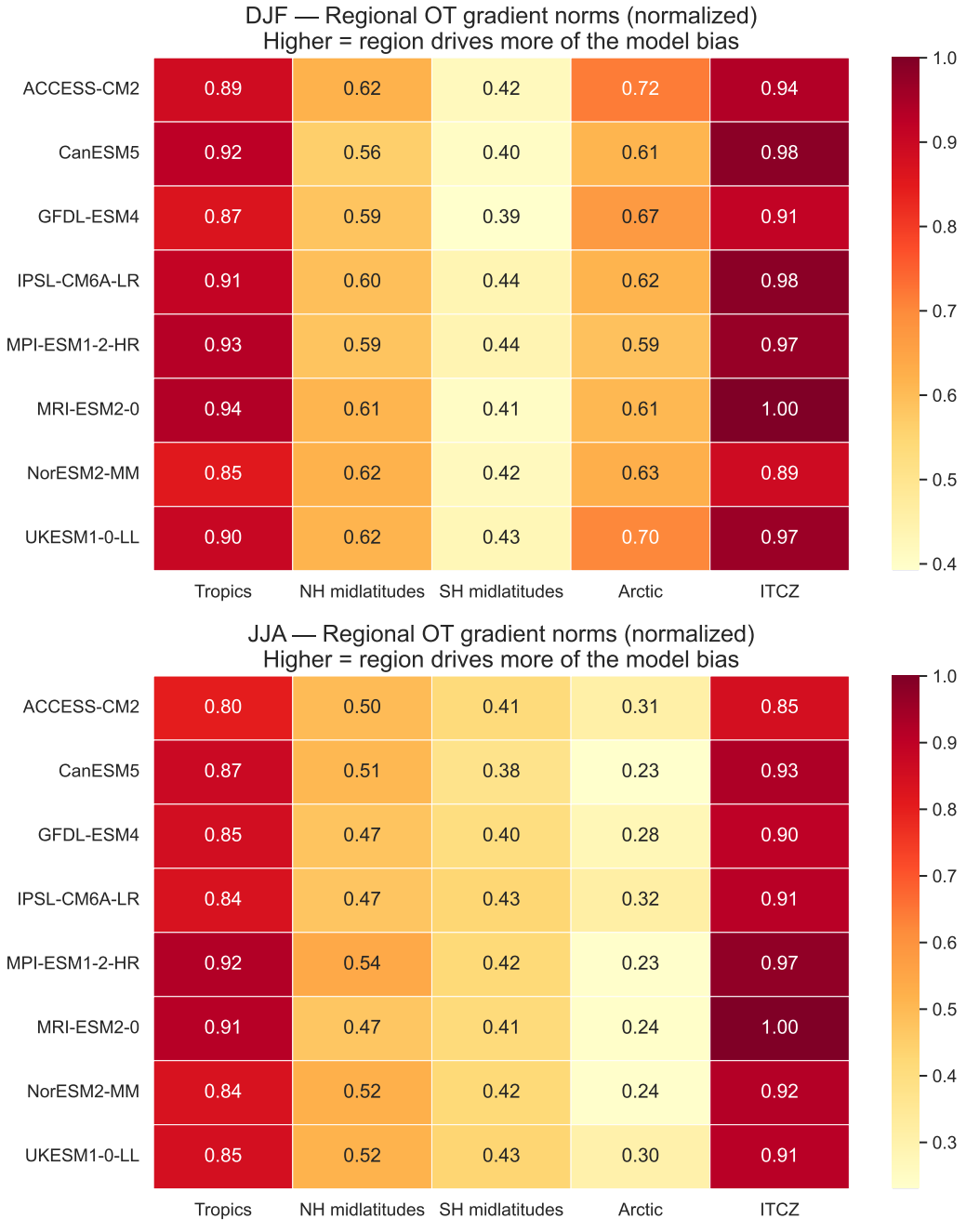


Figure 13: Regional OT gradient norms for eight CMIP6 models in DJF (up) and JJA (down), normalized by the global maximum across all models, regions, and seasons. Higher values indicate that a region contributes more to the total Sinkhorn divergence from ERA5. Tropical and ITCZ biases dominate in both seasons with consistent inter-model spread. The Arctic column is strongly elevated in DJF relative to JJA, while SH midlatitudes remain uniformly low across all models and seasons.