# 1. Equations

Monge theory

$$\inf_{T_{\#}\mu=\nu} \int_{\mathcal{X}} c\left(x, T(x)\right) d\mu(x)$$

Cost matrix (cost function)

$$C(x, y) = \| W(x - y) \|_2^2$$

Kantorovich theory

$$W_p(\mu, \nu) = \left( \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} \| x - y \|^p d\pi(x, y) \right)^{1/p}$$

Sinkhorn Distance

$$SD_C^{\varepsilon}(\pi) = \langle \pi, C \rangle - \varepsilon H(\pi)$$

# FAQ

# 2. Forward

## 2D MT Forward

Q: How is our 2D forward problem formulated and discretized?

A: We solve the 2D Maxwell's equations in the frequency domain. The system is discretized using a staggered-grid finite difference method, leading to a complex sparse linear system. We handle both TE and TM modes.

## Boundary Condition

What boundary conditions are implemented?

Q: Left and right boundaries are from 1D layered-earth solutions. The top boundaries uses dirichelet condition.
A: Additionally, the bottom boundary incorporates a skin-depth extrapolation to approximate a half-space.

# 3. Grid

## Grid resolution

Q: What's the resolution of our grid setting?
A: For the Cascadia profile, the computational domain is discretized into a 90 by 240 grid.
Vertically, we use 90 layers, including 10 layers for the air and 80 layers for the subsurface. The grid extends to a maximum depth of 300 km using exponential spacing.
Horizontally, core region grid features a uniform resolution of 2.5 km. We added padding zones on both sides that extend out to nearly 1000 km to eliminate the edge reflection.
For the AMT profile,

# 4. Inversion

## Time & Efficiency

Q: How does the computational efficiency and inversion time of the OT scheme compare to the conventional MSE approach? Is the additional overhead significant for practical applications?
A： As illustrated in the performance profiling, the per-iteration time for the OT scheme is slightly higher than that of the MSE.
It is important to note that the primary bottleneck in MT inversion lies in the forward modeling and gradient computation. The OT-related calculation accounts for less than 10% of the total wall-clock time. More importantly, the superior convexity OT scheme makes the overall inversion process more efficient and robust for complex geological structures.

Q: Could you provide some specific metrics regarding the computational speed? For instance, what is the wall-clock time per iteration for different grid sizes and frequency sets?
A: Certainly. To evaluate the practical performance, we benchmarked two cases with different complexities.For the Commemi 2D-4 model, with a 40 by 40 grid and 30 frequencies, the time per iteration is approximately 1.54 seconds.For the AMT field data involving a larger 90 by 240 grid and 18 frequencies, the iteration time is around 5.90 seconds.The efficiency of our engine stems from the implementation of frequency-level parallelism. More importantly, we do not rely solely on naive automatic differentiation; instead, we have integrated PyTorch's Autograd with a custom-built Adjoint State Method. This hybrid approach allows us to leverage the flexibility of deep learning frameworks for OT misfit gradients while maintaining the memory efficiency and high performance of adjoint-state sensitivity analysis for large-scale sparse systems.

## Regularization Decay Machanism

Q: How is the regularization coefficient set?
A: We designed an adaptive regularization coefficient machanism. Specifically, we dynamically correlate the magnitude of this coefficient with the ratio of the gradients from the regularization term and the data term. Throughout the inversion, the algorithm automatically calculates and adjusts the regularization coefficient based on the gradient information of the current step, ensuring that the gradients generated by the regularization term and the data term are always

maintained at a similar order of magnitude. This design allows the system to balance the weight between model constraints and data fitting throughout the inversion, ensuring both stability in the early stages and precise data-driven fitting in the final model.

# 5. Noise & Cost Function

## Cost function with defined weighting matrix

Q: What is the role of the cost matrix in the optimal transport algorithm, and why is it necessary to apply weighting to it?
A: The cost matrix plays a central metric role in the optimal transport algorithm, defining the penalty incurred when moving a unit of mass between distributions, typically using the squared Euclidean distance in our study. This matrix allows the algorithm to quantify the geometric discrepancy between two point clouds. Weighting the cost matrix is essential for effectively integrating observational noise into the inversion framework. By assigning different quality weights based on noise standard, we emphasizes high-quality data. This enables the inversion algorithm to automatically identify and suppress the gradient contributions from these noisy points, ultimately enhancing the overall robustness of the inversion results.

# 6. For Real Data

## Data fitting

Q: How is the data fitting performance of your inversion?
A: In synthetic test, our algorithm(figure1) achieved high-precision data fitting(figure2). For field data, the majority of frequencies and stations show excellent fitting results. As illustrated in the figure, a small portion of the data exhibits a global vertical shift, yet the shape of the curves matches the observations very well, and the algorithm successfully recovers the underlying subsurface structures. This phenomenon occurs primarily at adjacent stations where the observed data differ significantly, leading us to suspect the presence of severe static shift effects. Notably， traditional method is highly sensitive to such global offsets and often generates misleading gradients, while our approach focuses on the geometric shape matching of point clouds. This allows the algorithm to capture the correct response trends despite static shifts, ensuring the geological reliability of the inversion.

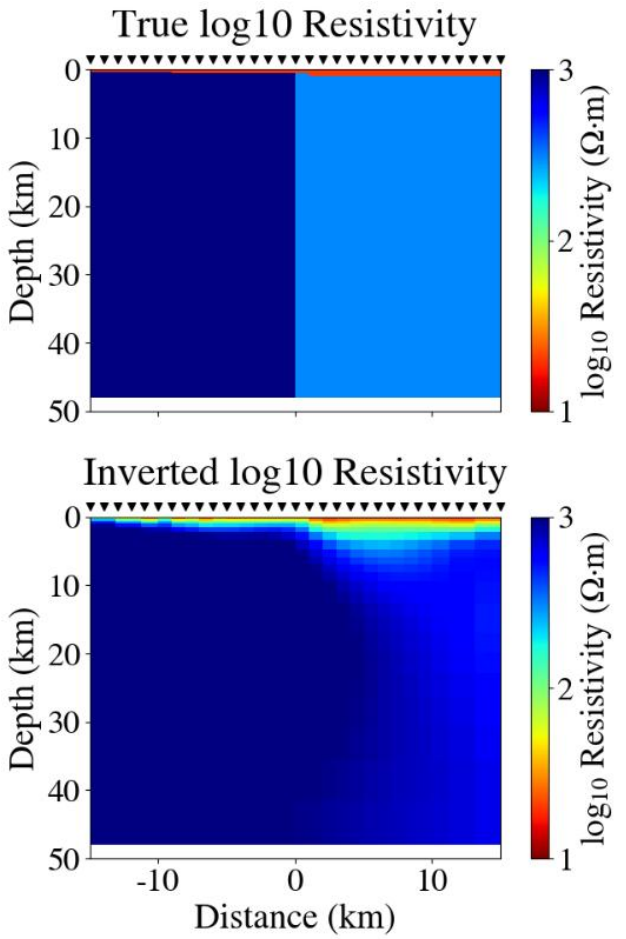


Figure 1 inversion redsult

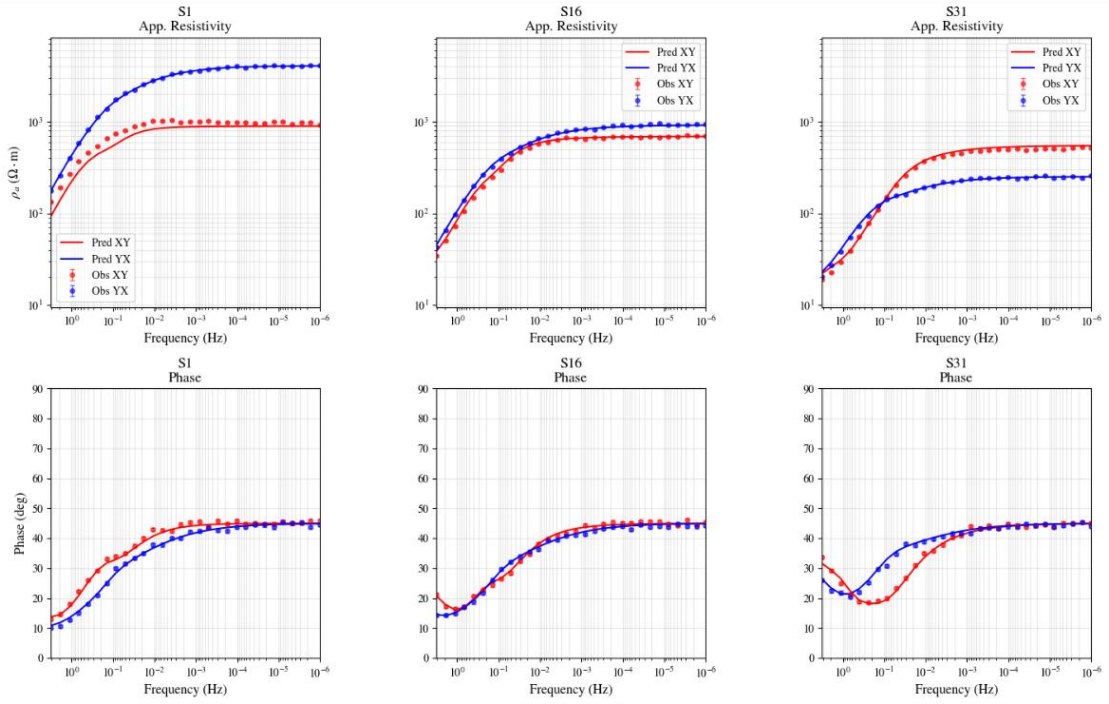


Figure 2 datafittng

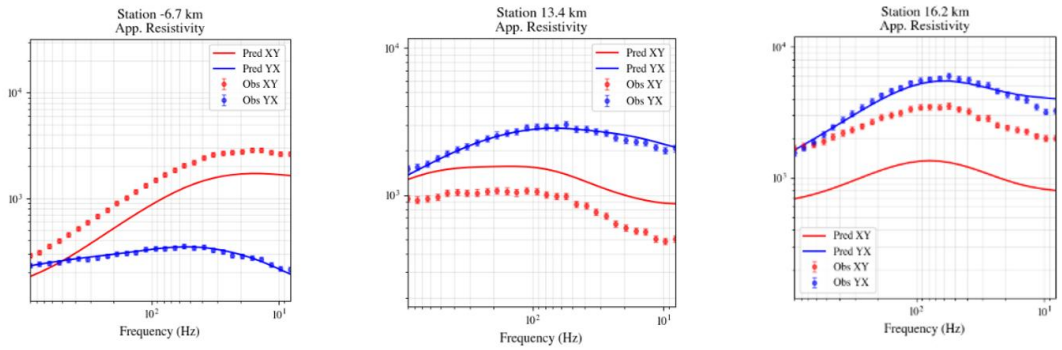


Figure 3 static shift

## Data Preparations

Q: What specific data preprocessing steps did you perform for the field measurements? 中文回
A: Before performing the inversion, we conducted rigorous quality control on the raw observational data. First, we applied magnetic declination correction to ensure the accuracy of the profile orientation. Second, we estimated the regional geological strike using phase tensor analysis and rotated the impedance tensor accordingly to align with the polarization directions required for 2D inversion. Additionally, we screened the stations to remove those with poor data quality and filtered out individual data points with strong 3D effects based on skewness indicators. These steps ensure that the data fed into our 2D inversion framework possess high reliability and physical consistency.

## SOURCE OF DATASET

AMT profile:
cascadia profile: Mocha & CAFE

## Recover New Subsurface Structure

Q: Regarding the field measurements such as the Cascadia profile, did your inversion results reveal any new geological structures?
A: When inverting the field data from the Cascadia profile, our method demonstrated superior structural resolution capabilities. First, in the upper left region of the profile, we clearly recovered the seawater layer, seafloor sediments, and the oceanic slab entirely driven by the data, without incorporating any prior model constraints. Second, beneath the oceanic slab, we successfully delineated a low-resistivity anomaly, which is highly consistent with the physical characteristics of a melt- and volatile-rich asthenospheric accumulation zone. Furthermore, conventional inversion methods typically have poor resolution for high-resistivity bodies and are prone to producing vertical stretching artifacts known as the hanging noodle effect in the deeper parts of the model. Our optimal transport algorithm not only significantly improves the resolving power for high-resistivity structures but also effectively mitigates these artifacts, resulting in a much clearer and more reasonable trend of the overall deep structural features.

## Definition Of Data Misfit

Q: How is the data misfit defined in your algorithm?

A: In our algorithm, data misfit is the measure of disagreement between predictions and observations: in mse mode it is the noise-weighted sum of squared residuals ( $\rho$ in log10 domain and phase in degree domain), while in 3dot/6dot mode it is the Sinkhorn optimal-transport distance between embedded point clouds, with RMS $\chi^2$ used as a unified statistical goodness-of-fit metric (ideally around 1).拟合优度

## A better performance of the measured data

Q: Do the field data show better inversion results for deep structures?

A: Yes. MSE solvers often suffer from blurred boundaries of deep anomalies or vertical stretching artifacts when dealing with deep structures due to the decreased resolution of electromagnetic data with increasing depth. In contrast, our optimal transport inversion effectively overcomes this by focusing on the feature matching of data point clouds based on global geometric shapes. In Cascadia subduction zone, not only are the high-resistivity features of the deep oceanic slab delineated more clearly and focused, but the geometry of the deep low-resistivity mantle corner flow beneath the slab is also more physically reasonable.

## The results of the cascadia profile inversion are different from those of MCGARY, Wannamaker and others

Q: Why do your inversion results differ from those published by McGary, Wannamaker, and others?

A: In this case, three factors matter. Compared to McGary's study, there are two main differences. First, McGary utilized prior oceanic-like model constraints, whereas our inversion is entirely data-driven. Second, we incorporated data from offshore stations, while their study relied exclusively on onshore constraints; the inclusion of offshore data significantly enhances the resolution of the forearc and shallow slab structures. ForWannamaker's research, he utilized a denser array of broadband station data, whereas we used a relatively smaller number of stations. Overall, the discrepancies mainly reflect differences in information content and modeling assumptions and our optimal transport algorithm successfully and robustly captured the key deep structural features.

# 7. Algorithm creation

## 3D OT

For $c \in \{\rho_{xy}, \phi_{xy}, \rho_{yx}, \phi_{yx}\}$

$$d_c^{obs} = [f, s, c]^{obs}$$
$$d_c^{pred} = [f, s, c]^{pred}$$
$$L_{OT}^{3D} = W_2\left(d_{\rho_{xy}}^{pred}, d_{\rho_{xy}}^{obs}\right) + W_2\left(d_{\phi_{xy}}^{pred}, d_{\phi_{xy}}^{obs}\right) + W_2\left(d_{\rho_{yz}}^{pred}, d_{\rho_{yz}}^{obs}\right) + W_2\left(d_{\phi_{yz}}^{pred}, d_{\phi_{yz}}^{obs}\right)$$

Why does the 3D OT inversion perform poorly compared to the 6D OT scheme in your study?
Thanks for question. THere are two primary reasons. First, 3DOT runs separate 3D OT fits on each MT data-component, so cross- component coupling is weak or absent. In contrast, the 6D OT embeds all four responses jointly in one 6D cloud,, enabling effective physical coupling. Also, 3DOT is more expensive because it solves one Sinkhorn OT problem per component

## The difference from the application in FWI

What are the differences and connections between your innovation and the current applications of Optimal Transport in Full Waveform Inversion?
While Optimal Transport shares a common goal in geophysical inversion by leveraging its global metric properties to overcome non-convexity, our approach differs significantly from its typical application in Full Waveform Inversion. In FWI, most OT implementations treat wave amplitudes as mass and travel time as the coordinate. Our innovation, however, tailors the OT framework to the multi-component nature of Magnetotelluric data. We embed frequency, station location, and various response components as coordinate dimensions, transforming the data into a point cloud of unit masses in a high-dimensional space. Since MT involves more physical quantities, we employ a 6D Optimal Transport scheme. This allows the algorithm to simultaneously capture geometric correlations across frequency, space, and physical parameters, rather than just matching a single waveform profile.

## The positivity of data

Is it necessary to ensure that all data values are positive when applying the optimal transport algorithm?
No, it is not necessary to ensure that the data are all positive. This is because, in our implementation, the various MT data components are embedded as coordinates in a high-dimensional space rather than being treated as the mass of the distribution. Traditional optimal transport requires the mass of the input distributions to be positive and the total mass between the source and target to be conserved. By transforming the measurements into a point cloud, we treat each data point as a discrete entity with an equal unit mass. This approach naturally satisfies the mathematical requirements of positive mass and mass conservation, eliminating the need for positivity-enforcing transformations such as logging or biasing, thereby

preserving the linear characteristics of the original data.

## The possibility of application in three dimensions

Is it possible to extend your algorithm to 3D magnetotelluric inversion?
Yes, our algorithm is highly scalable and possesses strong potential for extension to 3D inversion. Since our core approach involves embedding data components as coordinates in a high-dimensional space, extending it to 3D requirements is straightforward—we can easily incorporate additional data dimensions or coordinates to account for 3D observational features. This point-cloud-based strategy is independent of the model's dimensionality. Furthermore, because our framework integrates the efficient Sinkhorn algorithm with the adjoint state method, as long as the forward engine provides 3D responses, our optimal transport scheme can perform geometric shape matching in higher-dimensional spaces, enabling robust inversion of complex 3D geological structures.

## Preparatory work for becoming a probability space

What preliminary data preparations did your algorithm perform to ensure it satisfies the mathematical requirements of a probability space?
To satisfy the strict probability measure requirements of optimal transport theory, we initially transformed all observational data points into a discrete point cloud in a high-dimensional space. Subsequently, we applied a global normalization process to these point clouds, enforcing that the total mass of both the source and target distributions equals exactly one. Simultaneously, we assigned an equal unit mass to every single data point within both distributions. Through this point-cloud transformation and normalization approach, our data inherently satisfies the probability space requirements of non-negative and conserved total mass, perfectly aligning with the underlying mathematical assumptions of the optimal transport algorithm.

## unbalanced OT

$$L_{UOT}(a,b) = \min_{P\geq 0}\left[\langle P, C\rangle - \varepsilon H(P) + \lambda_1 D(P1 \parallel a) + \lambda_2 D(P^T 1 \parallel b)\right]$$

$a, b$: The empirical mass weight vectors of the source (predicted) and target (observed) point clouds.
$P$: The transport plan. $P1$ and $P^T 1$ are the actual marginal projections of the transport plan.
$\langle P, C\rangle$: The total transport cost, where $C$ is the cost matrix encoding the geometric discrepancy.
$-\varepsilon H(P)$: The entropy regularization term (Sinkhorn penalty) used to smooth the objective and enable efficient iterative solving.
$\mathcal{D}(\cdot \parallel \cdot)$: The divergence penalty (typically Kullback-Leibler divergence), which measures the mismatch between the transport plan’s marginals and the actual distributions.
$\lambda_1, \lambda_2$: The relaxation parameters (marginal penalty weights). As $\lambda \to \infty$, the divergence penalty becomes infinite, enforcing $P1 = a$ and $P^T 1 = b$, thereby recovering strictly balanced OT.

Q: Does your algorithm utilize balanced optimal transport? Have you considered applying unbalanced optimal transport?
A: Our algorithm utilizes strictly balanced optimal transport. By transforming the observational data into a discrete high-dimensional point cloud and assigning an equal unit mass to each data point, our approach naturally satisfies the mathematical requirements of balanced optimal transport, specifically that the mass of the source and target distributions must be non-negative and completely conserved. We did, in fact, experiment with unbalanced optimal transport during our preliminary tests, but the inversion results were unsatisfactory. Unbalanced optimal transport typically introduces relative entropy, such as KL divergence, to relax the marginal constraints, allowing for the creation or destruction of a certain amount of mass during the transport process. However, in our point-cloud-based physical context, this non-conservation of mass generates misleading physical gradients in the loss function, failing to effectively guide the subsurface resistivity model toward the true geological structures.

## AD’s Role

Q: What role does Automatic Differentiation play throughout the inversion process in your study?
A: Automatic differentiation serves as a core bridge in our entire inversion framework, accurately and automatically computing the gradients of the optimal transport loss. More importantly, we innovatively integrated automatic differentiation with the classic adjoint state method used in geophysics. The adjoint state method efficiently handles the sensitivity of the physical forward equations, while automatic differentiation manages the complex point cloud distance gradients. This combination not only ensures high computational efficiency and low memory footprint but also provides immense flexibility to the inversion framework, making it extremely easy to swap different loss functions or incorporate complex prior constraints, thereby laying the groundwork for integrating deep learning algorithms in the future.In other words, AD is what lets us treat OT transport regularization + geophysical forward physics as one differentiable pipeline, without manually deriving the sensitivity of the OT coupling to the model each time we change the embedding, weights, or OT hyperparameters.

## Is the transport restricted to fixed frequency and station pairs, or can mass be reallocated across frequencies / across stations?

Q: Will your algorithm involve mass transport across different frequencies or stations?
A: Yes, mass can be transported across different frequencies or neighboring stations. Conventional mean squared error can only perform isolated point-to-point numerical comparisons, which often generates misleading derivatives when individual data points are severely contaminated. In contrast, cross-frequency and cross-station mass transport allows the

algorithm to fit the overall distribution trend of the observational data from a global geometric perspective, rather than focusing on matching single isolated data points. This globally coordinated transport mechanism significantly enhances the stability and physical rationality of the inversion results.

Push：

Will this cross-station transportation leads to nonphysical problem?

No—Sinkhorn OT ‘transport’ is not a physical mass flux between stations. It is a geometric discrepancy measure between empirical distributions in an engineered feature space (frequency, site, responses). The Earth model is still constrained by the Maxwell- based forward operator, so we are not violating EM causality. OT does introduce nonlocal coupling in the data objective; if that coupling is too strong, we control it via cost weights on frequency/station dimensions and Sinkhorn regularization

## Loss Function

Q：Is the objective function used in your actual inversion process a strongly convex function?

A: Theoretically, the sinkhorn divergence component derived solely from the data terms is indeed characterized by strong convexity. However, in our practical inversion framework, we incorporate a model roughness regularization term to ensure physical consistency, which could theoretically influence the convexity of the data-driven term. To address this, we implemented an adaptive regularization strategy. During the initial stages, the regularization term stabilizes the process by constraining the model space. As iterations progress, the coefficient adaptively decreases, allowing the data term with its superior convexity to dominate the objective function. This approach leverages the global convergence advantages of optimal transport while significantly reducing the dependency on the initial model, ensuring that the final solution reliably converges toward the global minimum.

## OT’s necessity

Q：While OT is used in FWI primarily to address the cycle-skipping problem, such a phenomenon does not exist in MT. What then is the necessity of introducing OT into MT inversion?

A： It is true that MT inversion does not suffer from the cycle-skipping issues found in FWI; however, it still faces significant challenges regarding non-uniqueness and local minima. The necessity of introducing OT into MT lies first in leveraging its superior mathematical convexity, which effectively expands the convergence radius, reduces the dependency on the initial model, and enhances the algorithm's ability to converge under complex scenarios. Furthermore, our innovation involves embedding MT frequencies, station locations, and multiple physical components into a unified high-dimensional space for point-cloud matching. This approach breaks the limitations of conventional methods that fit components in isolation, greatly strengthening the inherent physical correlations between different data dimensions. This allows for the extraction of more robust geological features from complex observations.

## Geomloss Library

Q:  Why did you choose to use the Geomloss library and the PyTorch framework for your implementation?
A: The decision to use PyTorch and Geomloss was primarily driven by considerations of algorithmic flexibility and computational efficiency. First, the gradient calculation for Optimal Transport is extremely complex, involving implicit differentiation of the transport plan. The Geomloss library is perfectly compatible with PyTorch's automatic differentiation framework, allowing us to leverage its highly optimized Sinkhorn iterations to automatically obtain the gradients of the OT loss function with respect to model parameters. Second, PyTorch provides convenient vectorized operations and robust computational graph management. When combined with our custom adjoint state method, it significantly simplifies the programming of complex mathematical logic, such as high-dimensional point cloud matching, while maintaining physical accuracy. This integration not only shortens the development cycle but also establishes a foundation for future GPU acceleration.
Geometric Loss functions between sampled measures, images and volumes — GeomLoss