Extracting latent variables from forecast ensembles and advancements in similarity metric utilizing optimal transport

S. Nishizawa[1]

[1]RIKEN Center for Computational Science

Key Points:

- Novel method simplifies multiple spatial data, revealing hidden information efficiently for understanding probability distribution.
- Technique extracts essential similarities and differences in sparse distributions, aiding interpretation for improved analysis.
- Approach is adaptable to different data types, making it promising for diverse scientific fields.

Corresponding author: Seiya Nishizawa, `s-nishizawa@riken.jp`

Abstract

This study presents a novel methodology for extracting latent variables from high-dimensional sparse data, particularly emphasizing spatial distributions such as precipitation distribution. This approach utilizes multidimensional scaling (MDS) with a distance matrix derived from a new similarity metric, the Unbalanced Optimal Transport Score (UOTS). UOTS effectively captures discrepancies in spatial distributions while preserving physical units. This similar to mean absolute error however it considers location errors, providing robust measures crucial for understanding differences between observations, forecasts, and ensembles. Density estimation of these latent variables enhances the analytical utility, quantifying ensemble characteristics. The adaptability of the method to spatiotemporal data and its ability to handle errors suggest its potential as a promising tool for diverse research applications beyond spatial analysis in meteorology.

Plain Language Summary

This study introduces a new method to understand weather patterns by simplifying complex data. A mathematical technique was developed to efficiently identify hidden information from weather patterns. This assists meteorologists to understand the weather with greater accuracy. This method simplifies weather data by highlighting the essential similarities and differences between weather forecasts. This makes it easier for scientists to interpret and use the resultant data effectively. This study offers a new and efficient way to make sense of vast weather data, benefiting meteorological research, and potentially improving weather forecasting. This technique contributes to the meteorological field, in addition it also contributes to various fields with sparse distribution data.

1 Introduction

Probabilistic forecasts play a pivotal role in systems characterized by chaotic or stochastic behavior, such as weather forecasting. Ensemble simulations are commonly employed to estimate the probability distributions of future states. However, evaluating the predictive distribution in such multivariate, high-dimensional systems poses challenges, for instance in considering spatially distributed phenomena.

While univariate cases allow straightforward distribution definitions based on ensemble member results, multivariate cases, particularly in high-dimensional systems such as weather forecasting, face the "curse of dimensionality", Representing joint distribution which matches the state vector's dimensionality becomes infeasible owing to this issue, which influences accurate probability estimations.

Current discussions often focus on one-dimensional distributions, often considering points individually (e.g., grid points) or single statistical quantities, such as spatial averages. However, this point-wise approach could overlook crucial spatial patterns, especially in sparse quantities such as precipitation, leading to an overestimation of discrepancies between states, particularly in high-resolution simulation results (Gilleland et al., 2009).

Several systems operate within small embedded manifolds of lower dimensions, known as ranks. In addition, statistical methods often represent the observed variables in the original high-dimensional space as the outcomes of a mathematical model, with independent variables termed latent variables. The dimensionality of these latent variables can be significantly smaller than that of real variables. Herein lies the potential advantage of evaluating distributions in a lower-dimensional space, offering effective dimension-reduction opportunities.

Principal component analysis is widely used for dimensionality reduction, however it has limitations in nonlinear systems (for example, Nishizawa & Yoden, 2004). The vari-

ational autoencoder (Kingma & Welling, 2013) has displayed promise however it faces challenges in predictive forecast problems owing to limited ensemble sizes for training. It should be important to note that in this approach, the latent variables are often normalized and this normalization leads to the loss of information regarding the magnitude of the absolute values in the original physical space, preventing the spread of vectors corresponding to ensemble members from serving as an indicator of uncertainty. Moreover, this approach requires prior knowledge of the effective dimensions to effectively extract physically meaningful latent vectors in a smaller dimensional space.

In this study, a novel methodological approach for extracting meaningful latent variables from high-dimensional ensembles is proposed. These latent variables, which reside in a reduced space, aim to effectively represent the state of the system without prior knowledge of its effective dimensions. By preserving the physical units in distance measurements, this approach captures the essential distribution of the ensemble data. Through synthetic ensemble experiments, the effectiveness of this approach is demonstrated in extracting meaningful latent vectors.

2 Methods

2.1 Extracting Latent Variables

In this subsection, the proposed approach to extract latent variables from high-dimensional ensembles is described. The methodological approach was divided into three steps.

1. Calculation of a similarity metric for all pairs of ensemble members and observations.
2. Construction of a distance matrix from the calculated similarities.
3. Extraction of latent variables in a low-dimensional space from the distance matrix.

2.1.1 Metric for Similarity

Assessing the similarity between spatial distributions requires robust metrics that capture various discrepancies, including amplitude, location, area, and shape differences. In this study, various metrics were employed to measure the similarity between spatial distributions. Table 1 summarizes the metrics used in this study, each addressing the distinct facets of the discrepancies. These metrics include traditional metrics, such as the mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient (CORR). In addition, these include scores considering event-based dichotomous variables, such as the fraction skill score (FSS; N. M. Roberts & Lean, 2008; N. Roberts, 2008), equitable threat score (ETS; Gilbert, 1884), and frequency bias (FB; for example, Wilks, 2006). These were calculated point-wise, with the exception of FSS and FB, which are known to overestimate small-scale discrepancies. Among them, FSS is a score which considers spatial displacement and is widely used for high-resolution simulations. However, as it is based on a categorized quantity, it does not consider amplitude differences. When considering scores with different thresholds to simultaneously determine the event, the amplitude difference may be interpreted implicitly. In cases with a large number of samples, the interpretation of multiple scores may require complex and difficult considerations. For several purposes, such as probability distribution, a comprehensive single score is preferred. Furthermore, it is a dimensionless quantity. Thus, it is inappropriate to evacuate the spread of the distribution in the extracted latent variable space from FSS.

The Displacement and amplitude score (DAS; Keil & Craig, 2009) is a combination of displacement and amplitude differences. It contains more information than the traditional scores. However, there are several arbitrary definitions and computational

Table 1. Metrics for similarity used in this study

Abbreviation	Name	Distance
UOTS	Unbalanced optimal transport score	$UOTS$
DAS	Displacement and amplitude score	DAS
FSS	Fractions skill score	$1 - FSS$
MAE	Mean absolute error	MAE
RMSE	Root mean squared error	$RMSE$
CORR	Pearson correlation coefficient	$\sqrt{2(1 - CORR)}$
ETS	Equitable threat score	$1 - ETS$
FB	Frequency bias	$\lvert \log(FB) \rvert$

procedures. Keil and Craig (2007) showed that $D_{\max}$, which is the maximum search distance, has a great decisive impact on the result. It can only take discontinuous values: $D_{\max}$ is proportional to a power of two. Therefore, it may be difficult to choose an appropriate value based on physical considerations owing to its discontinuous constraint. They suggested that other parameters had a minor impact. However, non-negligible arbitrariness which they did not discuss exists. The score was defined such that the amplitude difference between one distribution and the morphed distribution of the other becomes the smallest; however, no condition was provided for the flow. In general, many possible flows can achieve the smallest amplitude difference. Thus, there are many possibilities for displacement, and the total score depends on the displacement. Another arbitrary factor is the difference in weight between the displacement and amplitude. This score is a combination of these two differences. As they have different units, the differences are normalized or nondimensionalized. The normalization factors are determined such that the two terms have equal weights; however, there is no justification for the weights to be equal. In addition, there is considerable arbitrariness in its computational procedure, resulting in a variation in the score. In fact, this study's implementation of computing the DAS results in a non-negligible difference in the obtained score compared to (Keil & Craig, 2009) for the same distributions owing to the undocumented details in the procedure. Another critical issue is that the procedure does not consider mass conservation during morphing.

Of significant note is the introduction of the Unbalanced Optimal Transport Score (UOTS) as a novel similarity metric. Designed specifically to evaluate spatial distribution discrepancies, the UOTS considers both amplitude and location differences in a unified manner, as does the DAS. Unlike DAS, the UOTS minimizes arbitrariness in its mathematical definition and offers a clearer physical interpretation, particularly regarding its hyperparameters L and q. In addition, the two terms of the displacement and amplitude differences have the same units and can be compared directly. Therefore, nondimensionalization does not need to be combined into a single score. The UOTS is a more straightforward score that considers both displacement and amplitude differences than DAS. The UOTS also has the same units as the original quantity, which facilitates physical interpretations.

2.1.2 Unbalanced Optimal Transport Score

The UOTS proposed in this study serves as a novel similarity metric tailored to assess spatial distribution discrepancies. The UOTS is defined as follows:

$$UOTS = \frac{1}{N} \min_{\gamma \in \mathbb{R}^{N^2}_{\geq 0}} \left\{ \sum_{i,j} \gamma_{ij} \left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2}{L} \right)^q + \frac{1}{2} \left(\|\gamma \mathbf{1} - \phi_1\|_1 + \|\gamma^T \mathbf{1} - \phi_2\|_1 \right) \right\}, \quad (1)$$

where $\mathbf{x}_i$ represents the location of the point i, $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ are mass distribution in the two distributions which are to be compared, and γ is the transport matrix, whose element γ_{ij} represents the mass transported from $\mathbf{x}_i$ to $\mathbf{x}_j$. $\mathbf{1}$ is a vector whose elements are all unity, and $\|\bullet\|_p$ represents the L^p norm. N is the vector length, i.e., $i, j = 1, \cdots, N$. The score is divided by N, however the number of nonzero elements can be used instead of N, depending on the purpose.

Defined as an optimization issue, the UOTS captures both amplitude and location differences, resembling the mean absolute error (L^1 norm) when considering spatial displacement. The first term in the brackets on the right-hand side penalizes location errors, whereas the second term represents the mean absolute error after the correction of location errors. $\gamma\mathbf{1}$ and $\gamma^T\mathbf{1}$ denote the mass distributions after transportation of ϕ_1 and ϕ_2, respectively.

Its formulation involves the hyperparameters L and q, which define the distance for identifying similar phenomena and the cost of transport per mass, respectively. The parameter L determines the distance threshold for identifying similar phenomena. Patterns exceeding this threshold are considered different. For the i and j index pairs, where $\|\mathbf{x}_i - \mathbf{x}_j\|_2 > L$, the optimal value of γ_{ij} must be zero; otherwise, the first term representing the transport cost outweighs the second term representing the amplitude difference. Larger q values downplayed the location difference, making the score more tolerant to small displacement errors.

The $UOTS$ multiplied by $L^q N$ is recognized as the optimal partial transport (Caffarelli & McCann, 2010; Chizat et al., 2018; Figalli, 2010), flat metric (Peyré et al., 2019) and Kantrovidge-Rubinshutain distance (Hanin, 1992; Lellmann et al., 2014). The minimization problem for this optimization can be solved effectively by using the Sinkhorm algorithm (Cuturi, 2013) with a reservoir of dustbin points by incorporating a regularization term $\lambda\Omega(\gamma)$. Here, Ω and λ represent the entropy regularization function and its coefficient, respectively, and $\Omega(\gamma) = \sum_{i,j} \gamma_{ij} \log(\gamma_{ij})$. In this study, the parameter λ was fine-tuned to the smallest possible value without causing computational divergence.

The UOTS introduces a novel approach to evaluate spatial distribution patterns, providing a robust means of quantifying similarities between spatial distributions.

2.1.3 Dimension Reduction and Extraction of Latent Variables

Dimensionality-reduction techniques have been employed to capture essential features and reduce high-dimensional ensemble data to a more manageable form. These techniques assist in extracting the underlying latent variables inherent to the data.

Before extracting the latent variables in a reduced space, a distance matrix was constructed from the similarity metric between the ensemble members and the observational data. In the process of constructing the distance matrix, it is crucial to transform metrics into values resembling distances that signify zero for identical distributions, nonnegatives, and symmetry, as detailed in Table 1.

In this study, multidimensional scaling (MDS; for example, Cox & Cox, 2000), specifically classical MDS or principal coordinate analysis, is utilized to construct a Euclidean space, where the distances between samples correspond to the given distance matrix. This method allows for the extraction of state vectors in Euclidean space, revealing the relative importance of each coordinate and the number of effective dimensions based on stress functions. The stress function S in MDS is computed as follows:

$$S = \sqrt{\frac{\sum_{m_i<m_j}(D_{m_i m_j} - d_{m_i m_j})^2}{\sum_{m_i<m_j} D_{m_i m_j}}}, \tag{2}$$

where $D_{m_i m_j}$ represents the distance in the reduced space, and $d_{m_i m_j}$ is the distance derived from the similarity metric in the original space between members m_i and m_j.

MDS operates as a linear procedure, and a nonlinear dimension reduction technique, Isomap (Tenenbaum et al., 2000), is also employed. Isomap extends MDS by capturing nonlinear manifolds embedded within the original space. By employing geodesic distance with a neighborhood graph, Isomap can be applied to complex data structures beyond linear representations. Employing Isomap The influence of the linear limitation of MDS on the extracted state vectors was examined by employing Isomap. While MDS and Isomap, the distance in the low-dimensiotal space is kept to have the same units as that of the similarity metric, other dimension reduction methods, such as locally linear embedding (LLE; Roweis & Saul, 2000), t-distributed stochastic neighbor embedding (t-SEN; Van der Maaten & Hinton, 2008), uniform manifold approximation and projection (UMAP; McInnes et al., 2018), and (DensMAP; Narayan et al., 2021), do not since they reconstruct low-dimensional variables based on weights or probability corresponding to other points.

The state vector obtained within the low-dimensional space through dimension reduction serves as an estimate of the latent variables. The validity of these latent variables significantly depends on the definition of the similarity metric used.

2.2 Synthetic Data Experiment

The synthetic data experiment was designed following the methodology detailed in Ahijevych et al. (2009) to illustrate the characteristics of various similarity metrics for assessing spatial distributions. A prescribed geometric spatial distribution mimicking the precipitation distribution was utilized. This distribution is described as follows:

$$\phi(x,y) = \begin{cases} 0, & \left(\frac{x-x_1}{a}\right)^2 + \left(\frac{y-y_1}{b}\right)^2 \geq 1 \\ \Phi_1, & \left(\frac{x-x_1}{a}\right)^2 + \left(\frac{y-y_1}{b}\right)^2 < 1, \quad \left(\frac{x-x_2}{0.4a}\right)^2 + \left(\frac{y-y_1}{0.4b}\right)^2 \geq 1 \\ \Phi_2, & \left(\frac{x-x_2}{0.4a}\right)^2 + \left(\frac{y-y_1}{0.4b}\right)^2 < 1 \end{cases}, \tag{3}$$

where $x_2 = x_1 + 0.4a, \Phi_2 = 2\Phi_1$, and $x = i\Delta x$ and $y = j\Delta x$, with $i = 0, 1, \cdots, 601, j = 0, 1, \cdots, 501$ and $\Delta x = 4$ km.

Six spatial distributions (Fig. 1) were created, including one reference (observation) and five target patterns (forecasts). The parameters (x_1, a, b) for the reference, pattern 1, pattern 2, pattern 3, pattern 4, and pattern 5 are $(200\Delta x, 25\Delta x, 100\Delta x)$, $(250\Delta x, 25\Delta x, 100\Delta x)$, $(400\Delta x, 25\Delta x, 100\Delta x)$, $(325\Delta x, 100\Delta x, 100\Delta x)$, $(325\Delta x, 100\Delta x, 25\Delta x)$, and $(325\Delta x, 200\Delta x, 100\Delta x)$, respectively. In all distributions, $y_1 = 250\Delta x$ and $\Phi_1 = 12.7$ mm. These distributions were employed to assess the characteristics of the various similarity metrics.

Furthermore, in this study, this geometric distribution was extended to ensemble forecasts and multiple cases. The observations and ensemble members were generated using specific parameters to simulate diverse scenarios, resulting in 100 cases with 50 ensemble members each.

The parameters for the observations are $(x_1^{\text{obs}}, y_1^{\text{obs}}, a^{\text{obs}}, b^{\text{obs}}, \Phi_1^{\text{obs}}) = (300\Delta x, 250\Delta x, \sqrt{\frac{A}{\pi\alpha}}, \sqrt{\frac{A\alpha}{\pi}}, 2^{\epsilon_3/2})$, where A and α are the area and aspect ratios, respectively, and $(A, \alpha) = (2^{\epsilon_1/2}\pi a_0 b_0, 4^{\epsilon_2/2}\frac{b_0}{a_0})$. The constants were set as $a_0 = 25\Delta x$ and $b_0 = 100\Delta x$. ϵ are random numbers with a standard normal distribution.

The parameters for ensemble members are $(x_1^{\text{fcs}}, y_1^{\text{fcs}}, a^{\text{fcs}}, b^{\text{fcs}}, \Phi_1^{\text{fsc}}) = (x_1^{\text{obs}} + 50\Delta x\epsilon_4, y_1^{\text{obs}} + 50\Delta x\epsilon_5, \sqrt{\frac{A^{\text{fcs}}}{\pi\alpha^{\text{fcs}}}}, \sqrt{\frac{A^{\text{fcs}}\alpha^{\text{fcs}}}{\pi}}, 2^{\epsilon_8/2}\Phi_1^{\text{obs}})$, where, $(A^{\text{fcs}}, \alpha^{\text{fcs}}) = (2^{\epsilon_6/2}A^{\text{obs}}, 4^{\epsilon_7/2}\alpha^{\text{obs}})$.

3 Results

The characteristics of the various similarity metrics were examined using geometric spatial distributions. The experiment involved multiple metrics and the sweeping of their hyperparameters. L and q for UOTS were swept at 200 km, $\leq L \leq$ 800 km, and $q \in (1, 2)$. The FSS also had a hyperparameter W which represents the width of neighborhoods, and it was swept from 200 to 800 km. The parameters for DAS were set to $D_{\max} = 180$ and 360 km, and $I_0 = 15.4$ m, which were determined according to previous research (Ahijevych et al., 2009).

Figure 1 visually demonstrates the magnitude of various similarity metrics applied to the five distributions with respect to the reference. UOTS displayed consistent rankings across patterns, indicating stability against parameter changes, which is a favorable trait. Conversely, DAS and FSS exhibited high sensitivity to their parameters, signifying the necessity for careful parameter selection. Traditional scores, such as RMSE, MAE, CORR, ETS, and FB, as previously reported by (Ahijevych et al., 2009), showed limitations in distinguishing between patterns 1, 2, and 4. These outcomes emphasize the advantages of UOTS as a robust similarity metric.

To demonstrate the extraction of latent variables and advantages of the UOTS, a synthetic data ensemble experiment was conducted (Section 2.2). Figure 2 presents the distributions of the estimated latent variables in two-dimensional space. Although these distributions represent a single case, their qualitative characteristics are consistent across all the cases.

When utilizing UOTS, DAS, and FSS with a moderate W, the first and second coordinates appear to be independent. Conversely, in cases employing RMSE, MAE, ETS, CORR, FB, and FSS with small and large W, these coordinates exhibit a relationship. The FB and FSS with a large W are nearly one-dimensional, relying solely on the first coordinate. FB and FSS with large W depend solely on the area difference and disregard other errors, leading to a one-dimensional latent variable distribution. ETS, CORR, and FSS with small W were distributed in a two-dimensional space however exhibited a rather one-dimensional structure. The distributions using MAE and RMSE display intermediate characteristics between the two-dimensional independent structure (e.g., with UOTS) and one-dimensional structure (e.g., ETS). With the independent parameters given in the distribution generation, the two coordinates are anticipated to be independent if the latent variables are successfully extracted as coordinates. Furthermore, as the ensemble members were generated by adding a normal random number to the observation parameters, the observation state was expected to be located near the origin in the latent variable space. With UOTS, FSS with a medium W, and FB, the observation was located near the origin, as expected. However, the observations are not positioned near the origin for the other cases. From this perspective, UOTS and FSS with medium W emerged as favorable similarity metrics among those investigated.

Generally, the discrepancies between two spatial distributions can be categorized into four types: amplitude, location, area, and shape. In this experiment, the location error had two dimensions, and the latent variable was expected to be five dimensions. The effective dimensionality D of the estimated latent variables is explored using the stress function (Fig. 3). It is expected that the stress will decrease for $D \leq 5$, and remain constant for $D > 5$. UOTS with specific parameter configurations exhibited an effective dimensionality of five, aligned with the expectations: $L \geq 400$ and $q = 1$, and $L = 200$ and $q = 2$. However, some metrics displayed a continuous stress reduction even beyond five dimensions, suggesting an overestimation of dimensionality: UOTS with $L = 200$ and $q = 1$, DAS, RMSE, MAE, ETS, and CORR. With FSS of $W = 200$ and 400, it becomes constant at $D = 4$ and 3, respectively. The FSS does not consider the amplitude error is related to this underestimation of the dimensionality. Even for the FSS with $W = 800$ and FB, the stress was almost constant for all D, corresponding to a one-

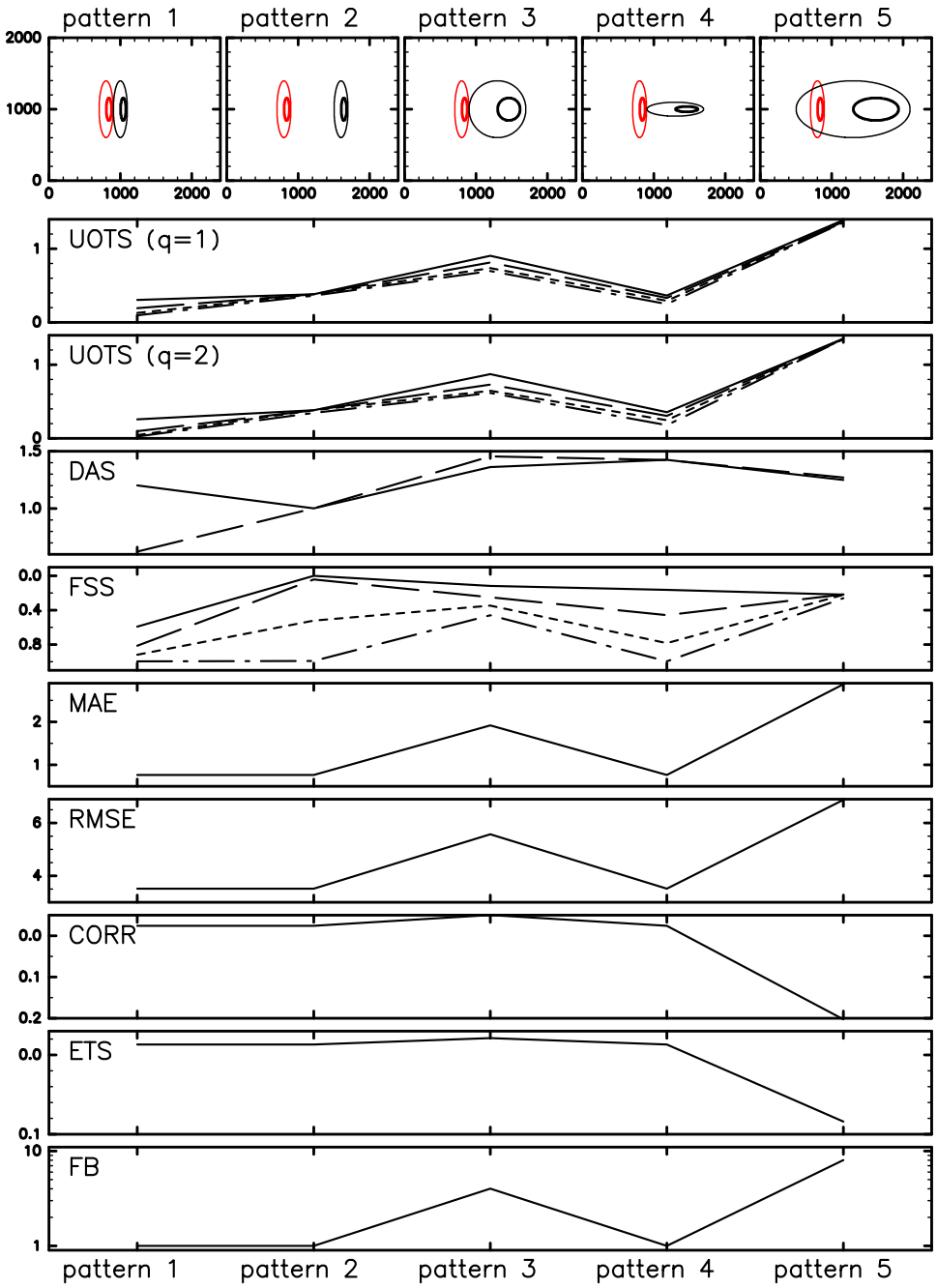

Figure 1. Spatial pattern of the synthetic geometric distributions. The top panels display the reference and five distributions. The red and black color indicates observation and forecasts, respectively. The thin and thick contours represents the area at which $\phi = 12.7$ and 25.4 mm, respectively. The lower panels show the magnitude of similarity metrics for the five distributions. Solid, dashed, dotted, and dash-dotted lines indicate $L = 200, 400, 600$, and 800 km for UOTS, $W = 200, 400, 600$, and 800 km for FSS, respectively. Solid and dashed lines indicate $D_{\max} = 180$, and 360 km for DAS, respectively.

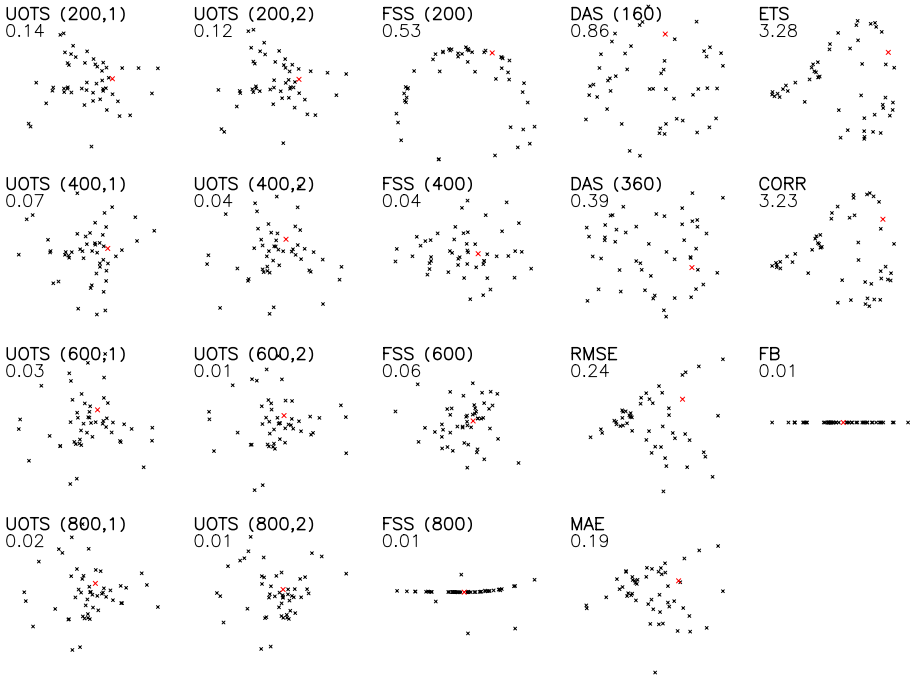

Figure 2. Spatial pattern of the estimated latent variables in the leading two-dimensional space. The black and red color indicates the ensemble member and observation, respectively. The numbers in parentheses represent L and q for UOTS, W for FSS, and $D_{\max}$ for DAS. The number under the metric name is the mean distance of the observation from the origin normalized by the standard deviation of the distance of members from the origin.

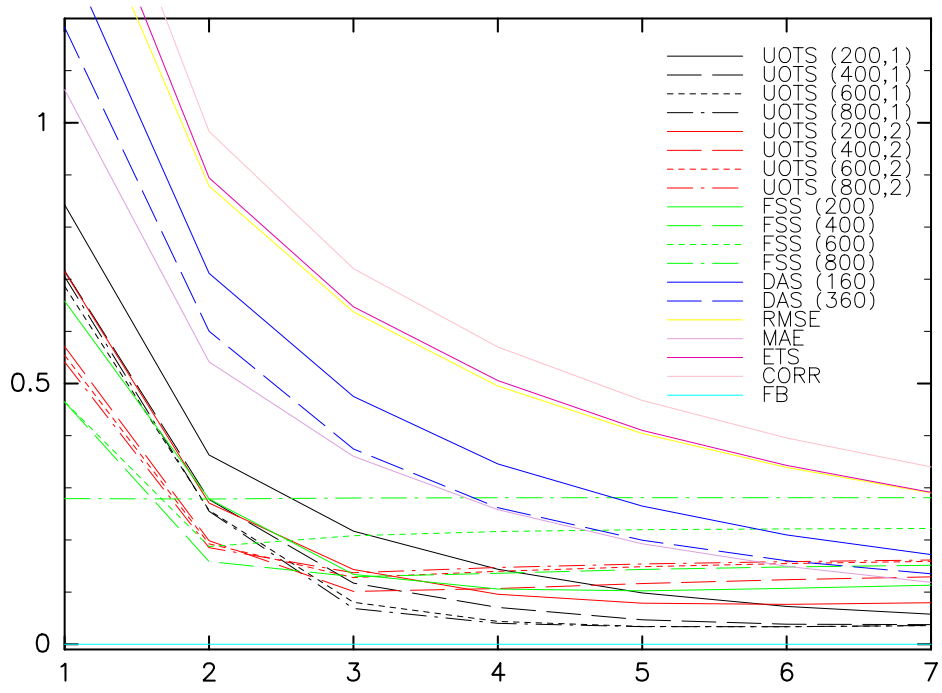

Figure 3. Dependency of the stress on the number of dimension D. Different color and line types indicate different metrics and parameters. The number in parentheses represent L and q for UOTS, W for FSS, and $D_{\max}$ for DAS.

dimensional structure. This indicates that significant information was being discarded. Conversely, the stress increases as D increases with UOTS with $L = 400, 600$, 800 and $q = 2$, and FSS with $W = 600$. This implies that these metrics are not appropriate for representing the Euclidean distance.

To investigate the relationship between the five parameters and the extracted coordinates, a correlation analysis was performed. In each case, the correlation coefficients between the parameters x, y, A, α and Φ and the extracted five leading variables were computed. As the order of the leading coordinates can vary depending on the case, the highest correlation coefficient among the five extracted coordinates was selected for each parameter. Figure 4 displays the correlation coefficient for each parameter and the ratio at which the correlation coefficient for the coordinates was the highest. The correlation coefficients for all five parameters were high for UOTS with a larger L, whereas some of the coefficients were small with other metrics. This confirms that UOTS successfully extracted the five latent variables. It should be noted that for UOTS with a larger L, A and Φ are almost equally represented by two coordinates. This is because the total mass (i.e., total precipitation), which is proportional to $A\Phi$, was extracted as a latent variable using the UOTS.

Overall, UOTS with larger L and $q = 1$ emerged as the most preferred similarity metrics among those investigated, providing insights into the latent variable distribution. Caution should be exercised regarding UOTS's sensitivity to an excessively large L, where penalties for transportation might be disregarded. This study primarily considers distributions with a single nonzero area, i.e., single phenomena, and practical scenarios involving multiple nonzero areas may require careful consideration of the appropriate L values to distinguish between them.

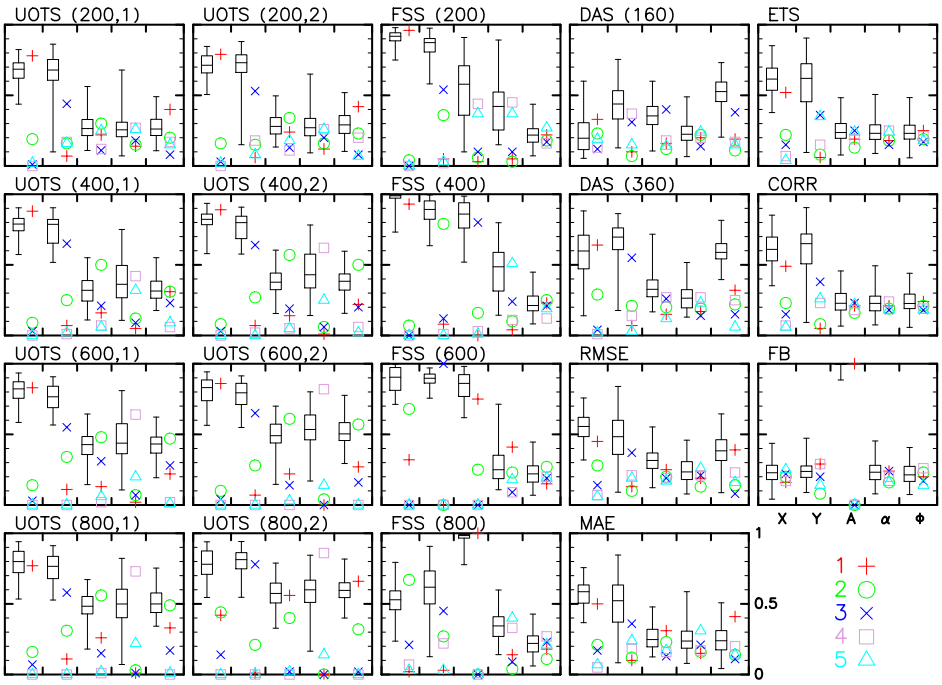

Figure 4. (Box-and-whisker plot) Correlation coefficient and (symbols) ratio of the number of coordinate with highest correlation. The box-and whisker plot represents, from the bottom, the minimum, the first quartile, the median, the third quartile, and the maximum. Red plus, green circle, blue x, purple square, and cyan triangle indicate the first, second, third, fourth, and fifth coordinates.

Furthermore, the linearity constraints inherited in MDS were considered. The distributions in two-dimensional space displays a one-dimensional structure with ETS, CORR, and FSS with $W = 200$. This can be attributed to the limitations of linearity inherent to MDS. To address this limitation, the nonlinear method Isomap was employed. However, the distributions in two-dimensional space obtained with Isomap are similar to those obtained using conventional MDS with ETS and CORR. For FSS with $W = 200$, although the shape changed significantly, it still exhibited a one-dimensional structure. This implies that the dimensionality constraint is inherent to the characteristics of the similarity metric.

4 Conclusions

This study introduces an innovative methodology aimed at deriving lower-dimensional latent variables from high-dimensional sparse data, primarily focusing on spatial distributions. The application of multidimensional scaling using a novel similarity metric, namely, the UOTS, has proven to be highly effective in extracting these latent variables. Notably, UOTS, similar to the mean absolute error however it considers location errors, provides a robust measure that retains physical meaning within its latent vectors.

The estimation of probability distributions from these latent variables using density estimation methods, such as kernel density estimation, offers substantial analytical advantages over the original high-dimensional space. This approach offers several potential advantages for various applications. For example, it enables the determination of the ensemble mean and spread while considering crucial factors such as location differences, which are vital in numerous meteorological applications. The ensemble mean can be established using the unbalanced optimal transport theory as the barycenter, whereas the ensemble spread can be derived from the squared sum of the eigenvalues obtained through multidimensional scaling. To evaluate the probability distribution and compare distributions in different cases, it is crucial that the Euqlidian distance in the latent variable space is almost identical to that in the original high-dimensional space. The UOTS has the same units as the original physical quantity and MDS preserves the units. Therefore, the method using the UOTS and MDS is preferable to consider the probability distribution in low-dimensional space.

Although the primary focus was on the spatial distributions, this method is adaptable to spatiotemporal distributions with minimal modifications. Incorporating factors, such as advection speed in the temporal direction, into the transport cost of UOTS allows for a seamless extension while maintaining the core methodology.

The efficacy of this methodology is underscored by its ability to handle discrepancies in spatial distributions by considering the amplitude, location, area, and shape errors. The UOTS with a larger L and $q = 1$ emerged as the most preferable similarity metric among those investigated to comprehend the distribution of latent variables.

The versatility of this approach can be extended to various meteorological applications. Moreover, this approach is not limited to meteorology as it is also applicable to various fields dealing with sparse spatiotemporal distributions. Its adaptability to diverse domains and robustness in handling errors makes it a promising tool across scientific disciplines.

Acknowledgments

This work was supported by JST [Moonshot R&D Program] Grant Number [JPMJMS2286]. The diagrams in this study were drawn using tools developed by the GFD-Dennou Club (http://www.gfd-dennou.org/).

Open Research Section

The programs for analysis visualization, and output data used in this study are available at Nishizawa (2023).

References

Ahijevych, D., Gilleland, E., Brown, B. G., & Ebert, E. E. (2009). Application of spatial verification methods to idealized and nwp-gridded precipitation forecasts. *Weather and Forecasting*, *24*(6), 1485–1497. doi: 10.1175/2009WAF2222298.1

Caffarelli, L. A., & McCann, R. J. (2010). Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, 673–730.

Chizat, L., Peyré, G., Schmitzer, B., & Vialard, F.-X. (2018). Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, *87*(314), 2563–2609. doi: 10.1090/mcom/3303

Cox, T. F., & Cox, M. A. (2000). *Multidimensional scaling*. CRC press.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26). Curran Associates, Inc.

Figalli, A. (2010). The optimal partial transport problem. *Archive for rational mechanics and analysis*, *195*(2), 533–560. doi: 10.1007/s00205-008-0212-7

Gilbert, G. K. (1884). Finley's tornado predictions. *American Meteorological Journal.*, *1*(5), 166.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., & Ebert, E. E. (2009). Intercomparison of spatial forecast verification methods. *Weather and forecasting*, *24*(5), 1416–1430. doi: 10.1175/2009WAF2222269.1

Hanin, L. G. (1992). Kantorovich-rubinstein norm and its application in the theory of lipschitz spaces. *Proceedings of the American Mathematical Society*, *115*(2), 345–352. doi: 10.1090/S0002-9939-1992-1097344-5

Keil, C., & Craig, G. C. (2007). A displacement-based error measure applied in a regional ensemble forecasting system. *Monthly Weather Review*, *135*(9), 3248–3259. doi: 10.1175/MWR3457.1

Keil, C., & Craig, G. C. (2009). A displacement and amplitude score employing an optical flow technique. *Weather and Forecasting*, *24*(5), 1297–1308. doi: 10.1175/2009WAF2222247.1

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. doi: 10.48550/arXiv.1312.6114

Lellmann, J., Lorenz, D. A., Schonlieb, C., & Valkonen, T. (2014). Imaging with kantorovich–rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, *7*(4), 2833–2859. doi: 10.1137/14097552

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. doi: 10.48550/arXiv.1802.03426

Narayan, A., Berger, B., & Cho, H. (2021). Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature biotechnology*, *39*(6), 765–774. doi: 10.1038/s41587-020-00801-7

Nishizawa, S. (2023). *Programs and data used for extracting latent variables from forecast ensembles and advancements in similarity metric utilizing optimal transport [DataSet].* Zenodo. `https://doi.org/10.5281/zenodo.10275595`.

Nishizawa, S., & Yoden, S. (2004). A parameter sweep experiment on topographic effects on the annular variability. *Journal of the Meteorological Society of Japan. Ser. II*, *82*(3), 879–893. doi: 10.2151/jmsj.2004.879

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With appli-

cations to data science. *Foundations and Trends® in Machine Learning*, *11*(5-6), 355–607. doi: 10.1561/2200000073

Roberts, N. (2008). Assessing the spatial and temporal variation in the skill of precipitation forecasts from an nwp model. *Meteorological Applications*, *15*(1), 163–169. doi: 10.1002/met.57

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. doi: 10.1175/2007MWR2123.1

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, *290*(5500), 2323–2326. doi: 10.1126/science.290.5500.2323

Tenenbaum, J. B., Silva, V. d., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, *290*(5500), 2319–2323. doi: 10.1126/science.290.5500.2319

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*, 2579–2605.

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (Vol. 91). Academic press.