Journal of the Meteorological Society of Japan

EARLY ONLINE RELEASE

This is a PDF of a manuscript that has been peer-reviewed and accepted for publication. As the article has not yet been formatted, copy edited or proofread, the final published version may be different from the early online release.

This pre-publication manuscript may be downloaded, distributed and used under the provisions of the Creative Commons Attribution 4.0 International (CC BY 4.0) license. It may be cited using the DOI below.

The DOI for this manuscript is
DOI:10.2151/jmsj.2021-047
J-STAGE Advance published date: April 29th, 2021
The final manuscript after publication will replace the preliminary version at the above DOI once it is available.

Statistical Intercomparison of Similarity Metrics in Sea Level Pressure Pattern Classification

Takuto SATO

Graduate School of Life and Environmental Sciences,
University of Tsukuba, Tsukuba, Japan

and

Hiroyuki KUSAKA[1]

Center for Computational Sciences,
University of Tsukuba, Tsukuba, Japan

January 25, 2021

---

1) Corresponding author: Hiroyuki Kusaka, Center for Computational Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577.
E-mail: kusaka@ccs.tsukuba.ac.jp.
Tel: +81-29-853-6481

Abstract

In this study, we compare the accuracy of five representative similarity metrics in extracting sea level pressure (SLP) patterns for accurate weather chart classification: correlation coefficient, Euclidean distance (EUC), S1-score (S1), structural similarity (SSIM), and average hash. We use a large amount of teacher data to statistically evaluate the accuracy of each metric. The evaluation results reveal that S1 and SSIM have the highest accuracy in terms of both average and maximum scores. Their accuracy does not change even when non-ideal data are used as the teacher data. In addition, S1 and SSIM can reproduce the subjective resemblance between two maps better than EUC. However, EUC reproduces the central position of the signal in a sample case. This study can serve as a reference for identifying the most useful similarity metric for the classification of SLP patterns, especially when using non-ideal teacher data.

Keywords similarity metrics; weather chart; classification

1. Introduction

Weather chart classification is widely used in meteorology and climatology. Various classification techniques have been proposed and their efficiency discussed (e.g., Huth et al., 2008). Classification techniques are either subjective or objective (computer-assisted). Computer-assisted classification methods are more objective than human perception and applicable to large datasets, and they are, therefore, popular in climatological analysis. Typical examples of such classification methods are clustering techniques, such as Ward's method, K-means clustering, and principal component analysis (PCA; e.g., Key and Crane, 1986; Cheng and Wallace, 1993; Hoffmann and Schkunzen, 2013, Kato et al., 2013; Miyasaka et al., 2020). Machine learning methods, including self-organizing maps (e.g., Hewiston and Crane, 2002; Cassano et al., 2006; Johnson et al., 2008; Ohba et al., 2016; Tamaki et al., 2018) and support vector machines (e.g., Kimura et al., 2009; Ortiz-García et al., 2014; Su et al., 2018), are now being used for classification or selection problems.

Similarity metrics are important in achieving accurate objective classifications. Hence, the accuracy of similarity metrics has been extensively discussed in meteorology and climatology (e.g., Gutzler and Shukla, 1984; Toth, 1991; Matulla et al., 2008; Tartaglione et al., 2009; Mo et al., 2014). Matulla et al. (2008) evaluated the accuracy of three conventional similarity metrics, i.e., correlation coefficient (COR), Euclidean norm, and S1-score (S1), with respect to the analog method, which is a forecasting method that uses historical data that have similarities with the target field. They found that the best choice of similarity metric

is dependent on the target variable of the analog method. Mo et al. (2014) investigated the ability of non-traditional similarity metrics, such as structural similarity (SSIM), to capture visual resemblance, finding that the best metric is dependent on the case (e.g., one-dimensional pattern or two-dimensional pattern). The findings of the aforementioned studies imply that it is difficult to identify a single similarity metric that can be applied to all problems and variables. In particular, it remains unclear which similarity metric is more effective when teacher data with noise are used.

In this study, we compare the accuracy of five similarity metrics in selecting sea level pressure (SLP) patterns: COR, Euclidean distance (EUC), S1, SSIM, and average hash (aHash). aHash is based on a different concept to the other metrics (detail information provided Section 2.3), and, therefore, evaluating its effectiveness can contribute to the development of rapid computer-assisted classification methods.

Another unique feature of the present paper is that we use a large teacher dataset both with and without noise (i.e., non-ideal teacher data and ideal teacher data, respectively). In actual classifications, it is not always possible to use ideal teacher data. For example, considering the selection of the cyclones over the Sea of Japan (CSoJs), ideal teacher data would only contain CSoJ information and would not consider other disturbances, whereas non-ideal teacher data would include other disturbances and cyclone shifting.

2. Method

2.1 Data and Target Pattern

We created SLP maps of the northwestern Pacific Ocean (120–170°E and 20–55°N) surrounding the Japanese Islands. These maps were recorded every six hours using JRA-55 reanalysis data (Kobayashi et al., 2015) from 2007 to 2016. The horizontal resolution is 1.25°. The target SLP pattern includes the cyclones over the Sea of Japan (hereafter referred to as CSoJ). Figure 1 shows an example of CSoJ. Since the said example is a typical SLP pattern that causes rainfall and snowfall in Japan, it is useful for discussing the accuracy of SLP map classification. CSoJs are frequently observed during spring and winter. However, we analyzed whole-year data because we aimed to select CSoJs by focusing on the similarities of SLP maps.

Prior to calculating the similarity between two SLP maps, a visual (subjective) labeling process was performed manually (Fig. 2) to identify the CSoJ and non-CSoJ maps. First, the authors selected the CSoJ candidates (839 maps) from all SLP maps (14,612 maps). The same number of non-CSoJ maps were randomly selected as non-CSoJ candidates from all SLP maps, except CSoJ candidates using random numbers. This process is called the first round of selection (Round 1 in Fig. 2). Next, we invited five experts to independently check the maps of CSoJ and non-CSoJ candidates. Among said maps, those that were recognized by more than three experts as CSoJ and non-CSoJ were labeled as CSoJ maps (328 maps) and non-CSoJ maps (788 maps), respectively (Round 2 in Fig. 2). Notably, the

CSoJ and non-CSoJ maps selected by five experts have relatively less subjective bias compared with the maps visually selected by the authors. Moreover, the five experts were unaware that the first round of selection was performed by the authors. Indeed, the number of maps selected in the second round was reduced compared with the first round. A total of 1,116 SLP maps (hereafter referred to as the labeled dataset) were selected for the experiment.

2.2 Experimental Design

We focused on evaluating the performance of similarity metrics in classifying SLP maps by executing the workflow (Fig. 3) outlined below for each similarity.

First, we selected distinct teacher CSoJ data from the labeled dataset (Step 1). Second, we calculated the similarity between the selected teacher data and all other data in the labeled dataset excluding teacher data (Step 2). Third, we sorted the labeled dataset in descending order of similarity (Step 3), and then we derived the ranking list from the sorted dataset. This procedure was repeated 328 times using all 328 sets of CSoJ data as the teacher data. We used a ranking list to calculate the selection rate (Eq. 1), which indicates the ability of similarity metrics to select SLP maps:

$$Selection\ rate = \frac{N^{p}_{selected}}{N_{total\ CoJ}}. \quad (1)$$

where $N_{total\ CoJ}$ denotes the number of CSoJs in the labeled dataset excluding teacher data, and $N^{p}_{extracted}$ denotes the number of SLP maps that have CSoJs in the top *p*-rank data in the ranking list, where *p* refers to the *p*-th and higher rank data used to calculate the selection rate. When the value of *p* is small and the selection rate is high, the CSoJs are typically concentrated in the upper part of the list, and it can be considered that the similarity metric performs effectively.

We evaluated the performance of five similarity metrics by counting the number of CSoJ maps belonging to the same group. For example, the resemble method was used in Huth (1996). Through statistically evaluating the result of all 328 sets of teacher CSoJ data, we determined which similarity metrics are useful for practical classification.

2.3 Similarity metrics

The similarity metrics we used were COR, EUC, S1, SSIM, and aHash. EUC, aHash, and S1 are sometimes considered as dissimilarity metrics. This is because said metrics sometimes produce large values, and the larger the value, the more dissimilar the pattern. However, with respect to the workflow employed in this study, “similarity” and “dissimilarity” assumed the same meaning when sorting was performed in descending or ascending order. Therefore, we refer to all five metrics as similarity metrics.

(a) *Correlation coefficient (Pearson’s correlation coefficient)*

We calculated the COR, which is frequently used in meteorology and climatology. We used the multidimensional vector, which is obtained by converting the two-dimensional SLP data into a one-dimensional vector. The COR *R*-value can be anywhere between -1 and 1; the closer it is to 1, the higher the similarity. R can be expressed by Eq. (2) (Taylor, 2001):

$$R = \frac{\frac{1}{N}\sum_N (A(i,j)-\bar{A})(B(i,j)-\bar{B})}{\sigma_A \sigma_B} \quad (2)$$

where A and B are two SLP maps, *N* is the total number of grid points, and σ is the standard deviation.

(b) *Euclidean distance*

We used the EUC method, which is also commonly applied in meteorology and climatology, as the similarity metric when we viewed two SLP patterns as a multidimensional vector. In this method, the similarity *D* can be expressed by Eq. (3).

$$D = \sqrt{\sum_{n_x}\sum_{n_y}\big(A(i,j) - B(i,j)\big)^2} \quad (3)$$

where n_x, and n_y denote the number of grid points in the *x*- and *y*-directions, respectively. The smaller the *D*, the higher the similarity. This metric is strongly affected by the spatial average value; thus, the resemble pattern is not recognized appropriately when spatial

165 average differs. Therefore, in this study, the spatial anomaly of SLP was used only for EUC
166 to avoid said bias. Notably, there is also a connection between the COR and EUC when
167 considering the spatial variance of two maps.
168
169 (c) *S1-score*
170 S1 was originally proposed by Teweles and Wobus (1954), and it is often used to verify
171 the predicted value of SLP or the geopotential height of numerical model outputs. This is
172 because it considers the horizontal gradients (Eq. 4). S1 takes a value of 0 and more; the
173 smaller the value, the greater the similarity between the two states:
174
175 $$S1 = 100\frac{\sum_N \frac{1}{N}\{|\partial_x D_i|+|\partial_y D_i|\}}{\sum_N \frac{1}{N}\{\max(|\partial_x F_i|, |\partial_x A_i|)+\max(|\partial_y F_i|, |\partial_y A_i|)\}} \quad (4)$$
176
177 where $\partial_x X = \frac{\partial X}{\partial x}$ and $\partial_y X = \frac{\partial X}{\partial y}$. F_i denotes the forecast value, A_i denotes the
178 reference value, and D_i denotes the difference between the forecast and reference values.
179 Therefore, in this study, F_i indicates the dataset, A_i indicates the teacher data, and D_i
180 indicates the difference between the teacher-data SLP map and maps in the dataset.
181
182 (d) *Structural similarity*
183 SSIM was originally proposed by Wang et al. (2004), and it uses brightness difference *l*(A,

B), contrast difference *c*(A, B), and structure difference *s*(A, B). These values are determined for a partial area of the image. The similarity of the entire image is obtained by averaging the results for the entire evaluation area. In the present study, we set a 5 × 5 grid box area as the evaluation area and scanned the whole SLP map. SSIM can be expressed by Eq. (5).

$$S(A,B) = l(A,B) \cdot c(A,B) \cdot s(A,B) = \left(\frac{2\mu_A\mu_B+C_1}{\mu_A^2+\mu_B^2+C_1}\right) \cdot \left(\frac{2\sigma_A\sigma_B+C_2}{\sigma_A^2+\sigma_B^2+C_2}\right) \cdot \left(\frac{\sigma_{AB}+C_3}{\sigma_A\sigma_B+C_3}\right) \quad (5)$$

where $C_1 = C_2 = C_3 = 0.0$, μ is the average value in the evaluation area, σ is the standard deviation in the evaluation area, and σ_{AB} is the covariance of *A* and *B*. *S* has a value of $-1 \leq S \leq 1$; the closer it is to 1, the higher the similarity.

(e) *Average Hash*

The aHash method is mainly used in image retrieval. It was introduced by Krawetz (2011) and used by Fei et al. (2015) and Haviana and Kurniadi (2016). A similar concept exists, called image hashing, which is used in the rapid retrieval of two-dimensional meteorological fields (Raoult et al., 2018). The resolution of the target image is reduced to a specific number of pixels, and the image is hashed into a bit string with 1 and 0. The criterion is the average value of the image; 1 is hashed when the value is higher than the average value, and 0 is hashed otherwise. In this method, the Hamming distance (total number of different bits) of these bit strings is used as the similarity metric (separation degree); the smaller the

Hamming distance, the higher the similarity (i.e., the same bit string is hashed). In this study, the resolution was reduced to 16 $\times$ 16 pixels and hashed into a 256-bit bit string.

Considering these features, EUC and COR emphasize statistical elements (i.e., differences and correlations), SSIM and aHash emphasize visual elements (i.e., visual or perceptual similarity), and S1 emphasizes meteorological elements (i.e., pressure gradient). An intercomparison of the performance of these metrics can reveal which is better is classifying SLP maps.

3. Results

Table 1 summarizes the mean, maximum, and minimum selection rate when $p = 327$, where 327 is the data number of CSoJ maps excluding teacher data. S1 has the highest score for both the mean and maximum selection rate. SSIM has almost the same value as S1 in the mean selection rate. The other three metrics have a low selection rate, and there are no large differences between them. The minimum selection rate exhibits a trend similar to that of the mean selection rate. Wilcoxon's rank sum test was independently conducted for the five metrics at a significance level of 5%. Significant differences were observed between aHash and the other four metrics.

Figure 4 shows the relationship between the number of selected data and the selection rate. When the graph is convex upward, the selection accuracy is high because the CSoJ data are selected even with a small p value. SSIM and S1 have many upward convex graphs and fewer downward convex graphs compared with the other metrics (Fig. 4). In addition, the mean selection rate of S1 and SSIM (thick black line in S1 and SSIM panel) for all teacher data cases has stronger upward convection than the other three metrics.

Unlike the other four similarity metrics, aHash generally reduced the resolution of images during preprocessing to remove high-frequency components. However, this reduction in resolution can affect accuracy. To investigate this effect, we compare the selection rate with the image resolutions of 41 $\times$ 29 (unreduced), 16 $\times$ 16, and 8 $\times$ 8, the mean selection rates of which are 39.3%, 40.4%, and 40.7%, respectively. A multi-group test at a

significance level of 5% (Wilcoxon rank sum test using the Bonferroni method for modification) was performed. The null hypothesis was not rejected.

aHash binarizes the SLP pattern for calculation, which means that the horizontal gradient information is lost. Therefore, the effect of aHash binarizing should be investigated. Hashing is effective in general images because it extracts feature information. However, in the case of SLP patterns, it is difficult to distinguish between the features and the background. Thus, important classification information may be lost, and the accuracy may be reduced. Accordingly, we compared the selection rates between aHash and EUC for a 16 $\times$ 16 image resolution. The mean selection rates for all 328 teacher datasets are 40.4% for aHash and 40.0% for EUC. The null hypothesis was not rejected based on the results of Wilcoxon's rank sum test at a significance level of 5%.

Figure 5 shows some selected examples when the map of 06 UTC on March 5, 2007, is used as teacher data. In this teacher data, the CSoJ center is located north of the Sea of Japan with no other notable signals. The map of 06 UTC on March 1, 2013, has a high rank in terms of all three similarity metrics (S1: sixth; EUC: third; SSIM: ninth), the map of 00 UTC on February 14, 2007, has a high rank in terms of S1 and SSIM (S1: 12^{th}; EUC: 84^{th}; SSIM: 15^{th}), and the map of 18 UTC on November 6, 2012, has a high rank in terms of EUC (S1: 82^{nd}; EUC: 14^{th}; SSIM: 76^{th}). Herein, the rank refers to the ranking in the sorted dataset, and the ranking list is shown in Figure 2. A small rank number means that the data are more similar to the teacher data on a certain similarity metric scale. The map of 06 UTC on March

1, 2013, represents the visual similarity well. Specifically, the CSoJ shape and center and the anti-cyclone appearance over the Pacific Ocean are similar. Alternatively, the map of 00 UTC on February 14, 2007, does not represent the visual resemblance well. Specifically, the CSoJ center is on the south side of the Sea of Japan. In the map of 18 UTC on November 6, 2012, the CSoJ center is on the north side of the Sea of Japan, but the signal is not clear. In addition, the CSoJ shapes and anti-cyclone are dissimilar to those of the teacher data.

4. Discussion

The intercomparison of five similarity metrics showed that S1 and SSIM have the highest accuracy in the mean and maximum selection rates. In addition, the performance of these two metrics hardly deteriorates. Thus, using S1 and SSIM for classification can further improve the accuracy of existing classification methods.

S1 includes the horizontal pressure gradient in the metric. In the case of SLP patterns, the difference between the values of surrounding grid points (i.e., pressure gradient information) is important for classification. The result of this study supports this perspective. However, it should be noted that S1 performs poorly in some cases (Toth, 1991). Thus, the accuracy of S1 needs to be discussed further.

SSIM also has a high score comparable to S1. The fact that SSIM has a higher score than EUC and COR in the two-dimensional images is consistent with the results obtained by Mo et al. (2014). SSIM comprehensively considers the difference of mean, contrast, and structure. These aspects explain the visual resemblance of two SLP patterns. Therefore, in this study, SSIM was used to select the resembling maps. Previous studies confirm that SSIM is more accurate than the mean squared error for general images (e.g., Wang and Bovik, 2002). The mean squared error produces results that are similar to EUC in the present study. These results indicate the difficulties in determining the metrics that can completely capture the visual resemblance.

We also consider the effect of binarization on selection accuracy. There are no differences

between the mean value of the EUC and aHash selection rates. Based on this result, binarization may not be effective for classification. As suggested by the result of S1, the horizontal gradient information is important in classifying SLP patterns, and SLP pattern binarization may reduce classification performance due to a lack of horizontal gradient representation. However, when the signal is simple and the magnitude of the gradient becomes less important, aHash may be more accurate than other methods. Therefore, it is necessary to examine the accuracy of various patterns.

Another important aspect is the elements that each similarity metric emphasizes. Discussing how these factors affect the visual similarity provides insight into how to improve classification accuracy.

Let us consider EUC, S1, and SSIM as examples. The patterns that rank high in all metrics are visually (and subjectively) similar (Fig. 5). Specifically, the CSoJ shape and center and the appearance of the anti-cyclone over the Pacific Ocean are similar. In contrast, in the example that ranks high for S1 and SSIM, the CSoJ center is slightly different, but the overall pattern is visually similar. In the example that ranks high for EUC, the overall CSoJ pattern and features are inconsistent with the teacher data. These results suggest that the reproducibility of visual similarity is dependent on the metric being used. Therefore, not only the selection rate but also the characteristics of each similarity metric should be considered.

Finally, we describe the limitations of this study. We used the SLP maps belonging to CSoJ patterns as the teacher data. However, as mentioned earlier, using teacher data that are

inconsistent with the CSoJ pattern may change the results. Such variations are caused by a difference in the signal–noise ratio of the teacher SLP pattern; if the pattern has a high signal–noise ratio, the accuracy is high even when using a method based on image simplification, such as aHash. Similarly, the accuracy can be improved through preprocessing methods, such as PCA, for noise removal and dimension reduction. The spatial resolution of the data is also important. When the resolution increases, the dimension of the vector to be calculated in EUC and COR increases. Thus, the degree of separation decreases.

5. Conclusion and remarks

In this study, we evaluated the accuracy of five similarity metrics using a large amount of teacher data: COR, EUC, S1, SSIM, and aHash. These metrics are important in the classification of SLP patterns. The evaluation results revealed that S1 performed the best in terms of mean accuracy. It also performed the best in terms of maximum accuracy for the experiments using all teacher data. The second-best metric was SSIM. In other words, these two metrics reasonably classified SLP maps, even when teacher data were used with high noise. This suggests that S1 and SSIM can perform better than the other similarity metrics in practical classification problems. The accuracy of aHash was comparable to that of EUC and COR, which can be attributed to the resolution reduction and binarization performed in aHash.

Based on these results, S1 and SSIM are the most effective methods for the classification of SLP patterns. Thus, similarity metrics that emphasized visual factors (e.g., SSIM) and meteorological factors (e.g., S1) were useful for extracting the targeted SLP patterns. These metrics were able to reproduce the visual resemblance better than EUC. However, EUC was able to extract the CSoJ central position.

In the future, similar experiments focusing on other typical SLP patterns should be conducted to obtain more robust results. In addition, noise reduction techniques, such as PCA or the use of filters, could be compared to improve classification and extraction.

Acknowledgments

This research was financially supported by the Environment Research and Technology Development Fund (JPMEERF20192005) of the Environmental Restoration and Conservation Agency of Japan. We would like to thank the five co-workers who assisted us in the corporate classification, as well as the five experts who participated in the classification test, of the SLP patterns. We would also like to also express our gratitude to Dr. Akifumi Nishi for his advice.

References

Cassano, E. N., A. H. Lynch, J. J. Cassano, and M. R. Koslow, 2006: Classification of synoptic patterns in the western Arctic associated with extreme events at Barrow, Alaska, USA. *Climate Res.*, **30**, 83–97.

Cheng, X., and J. M. Wallace, 1993: Cluster analysis of the Northern Hemisphere wintertime 500-hPa height field: Spatial patterns. *J. Atmos. Sci.*, **50**, 2674–2696.

Fei, M., J. Li, L. Shao, Z. Ju, and G. Ouyang, 2015: Robust visual tracking based on improved perceptural hashing for robot vision. *Intelligent Robotics and Applications*, H. Liu, N. Kubota, X. Zhu, R. Dillman (eds.), Vol. 9246, Springer, Cham, 331–340.

Gutzler, D. S., and J. Shukla, 1984: Analogs in the wintertime 500 mb height field. *J. Atmos. Sci.*, **41**, 177–189.

Haviana, S. F. C., and D. Kurniadi, 2016: Average hashing for perceptual image similarity in mobile phone application. *J. Telematics Inf.* **4**, 12–18.

Hewitson, B. C., and R. G. Crane, 2002: Self-organizing maps: Applications to synoptic climatology. *Climate Res.*, **22**, 13–26.

Hoffmann, P., and K. H. Schlünzen, 2013: Weather pattern classification to represent the urban heat island in present and future climate. *J. Appl. Meteor. Climatol.*, **52**, 2699–2714.

Huth, R., 1996: An intercomparison of computer-assisted circulation classification methods. *Int. J. Climatol.*, **16**, 893–922.

Huth, R., C. Beck, A. Philipp, M. Demuzere, Z. Ustrnul, M. Cahynová, J. Kyselý, and O. E. Tveito, 2008: Classifications of atmospheric circulation patterns: Recent advances and applications. *Ann. New York Acad. Sci.*, **1146**, 105–152.

Johnson, N. C., S. B. Feldstein, and B. Tremblay, 2008: The continuum of Northern Hemisphere teleconnection patterns and a description of the NAO shift with the use of self-organizing maps. *J. Climate*, **21**, 6354–6371.

Kato, H., Y. Nagano, and S. Tanaka, 2013: Objective classification of the sea level pressure distribution pattern in East Asia: Analysis of the cold half of the year. *Geogr. Rev. Japan Series A,* **86**, 95–114 (in Japanese with English abstract).

Key, J. R., and R. G. Crane, 1986: A comparison of synoptic classification schemes based on ‘objective’ procedures. *Int. J. Climatol*, **6**, 375–388.

Kimura, H., H. Kawashima, H. Kusaka, and H. Kitagawa, 2009: Detection of pressure patterns using support vector machine: Winter type pressure pattern. *Geogr. Rev. Japan Series A*, **82**, 323–331 (in Japanese with English abstract).

Kobayashi, S., Y. Ota, Y. Harada, A. Ebita, M. Moriya, H. Onoda, K. Onogi, H. Kamahori, C. Kobayashi, H. Endo, K. Miyaoka, and K. Takahashi, 2015: The JRA-55 reanalysis: General specifications and basic characteristics. *J. Meteor. Soc. Japan*, **93**, 5–48.

Krawetz, N., 2011: Looks like it. [Available at http://hackerfactor.com/blog/index.php%3F/archives/432-Looks-Like-It.html.]

Matulla, C., X. Zhang, X. L. Wang, J. Wang, E. Zorita, S. Wagner, and H. von Storch, 2008: Influence of similarity measures on the performance of the analog method for downscaling daily precipitation. *Climate Dyn.*, **30**, 133–144.

Miyasaka, T., H. Kawase, T. Nakaegawa, Y. Imada, and I. Takayabu, 2020: Future projections of heavy precipitation in Kanto and associated weather patterns using large ensemble high-resolution simulations. *SOLA*, **16**, 125–131.

Mo, R., C. Ye, and P. H. Whitfield, 2014: Application potential of four nontraditional similarity metrics in hydrometeorology. *J. Hydrometeor.*, **15**, 1862–1880.

Ohba, M., S. Kadokura, D. Nohara, and Y. Toyoda, 2016: Rainfall downscaling of weekly ensemble forecasts using self-organising maps. *Tellus A*, **68**, 29293, doi:10.3402/tellusa.v68.29293.

Ortiz-García, E. G., S. Salcedo-Sanz, and C. Casanova-Mateo, 2014: Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data. *Atmos. Res.*, **139**, 128–136.

Raoult, B., G. Di Fitta, F. Pappenberger, and B. Lawrence, 2018: Fast retrieval of weather analogues in a multi-petabytes archive using wavelet-based fingerprints. *International Conference of Computational Science*, Shi, Y., H. Fu, Y. Tian, V. V. Krzhizhanovskaya, M. H. Lees, J. Dongarra, and P. M. A. Sloot (eds.), Vol. 10861, Springer, 697–710.

Su, S.-H., J.-L. Chu, T.-S. Yo, and L.-Y. Lin, 2018: Identification of synoptic weather types over Taiwan area with multiple classifiers. *Atmos. Sci. Lett.*, **19**, e861, doi:10.1002/asl.861.

Tamaki, Y., M. Inatsu, D. Nguyan-Le, and T. J. Yamada, 2018: Heavy rainfall duration bias in dynamical downscaling and its related synoptic patterns in summertime Asian monsoon. *J. Appl. Meteor. Climatol.*, **57**, 1477–1496.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192.

Tartaglione, N., M. Maugeri, F. Dalan, M. Brunetti, T. Nanni, and A. Speranza, 2009: Searching for resemblance between large-scale sea level pressure patterns leading to “intense” precipitation events over Italy. *Theor. Appl. Climatol.*, **95**, 183–196.

Teweles Jr., S., and H. B. Wobus, 1954: Verification of prognostic charts. *Bull. Amer. Meteor. Soc.*, **35**, 455–463.

Toth, Z.,1991: Intercomparison of circulation similarity measures. *Mon. Wea. Rev.*, **119**, 55–64.

Wang, Z., and A. C. Bovik, 2002: A universal image quality index. *IEEE Signal Process. Lett.*, **9**, 81–84.

Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, 2004: Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, **13**, 600–612.

Table 1. Mean, maximum, and minimum selection rate (%) for each similarity metric. The selection rate is the ratio of CSoJ in the top 327 members (p=327) of the sorted data as shown in Eq. (1).

Metrics	Mean	Maximum	Minimum
COR	43.0	73.1	19.6
EUC	39.0	64.2	19.9
aHash	40.4	66.7	15.3
SSIM	56.6	77.7	26.6
S1	57.5	80.4	23.5

418 List of Figures

419

430

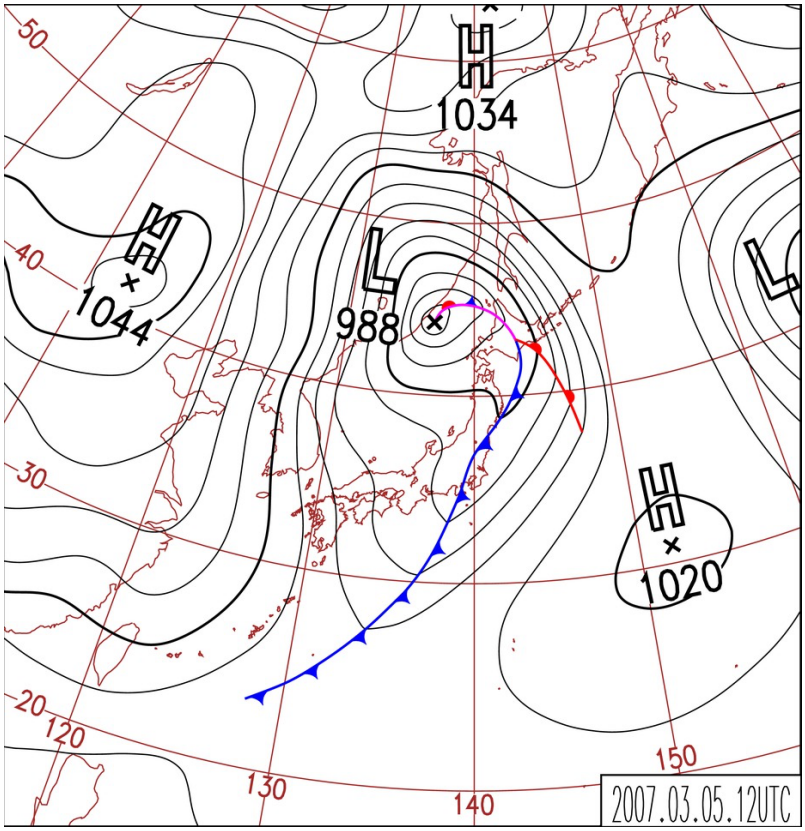


Fig.1 An example of a weather chart with a CSoJ pattern (12 UTC on March 5, 2007). (Created by National Institute of Informatics "Digital Typhoon" based on "Weather Charts" from Japan Meteorlogical Agency)

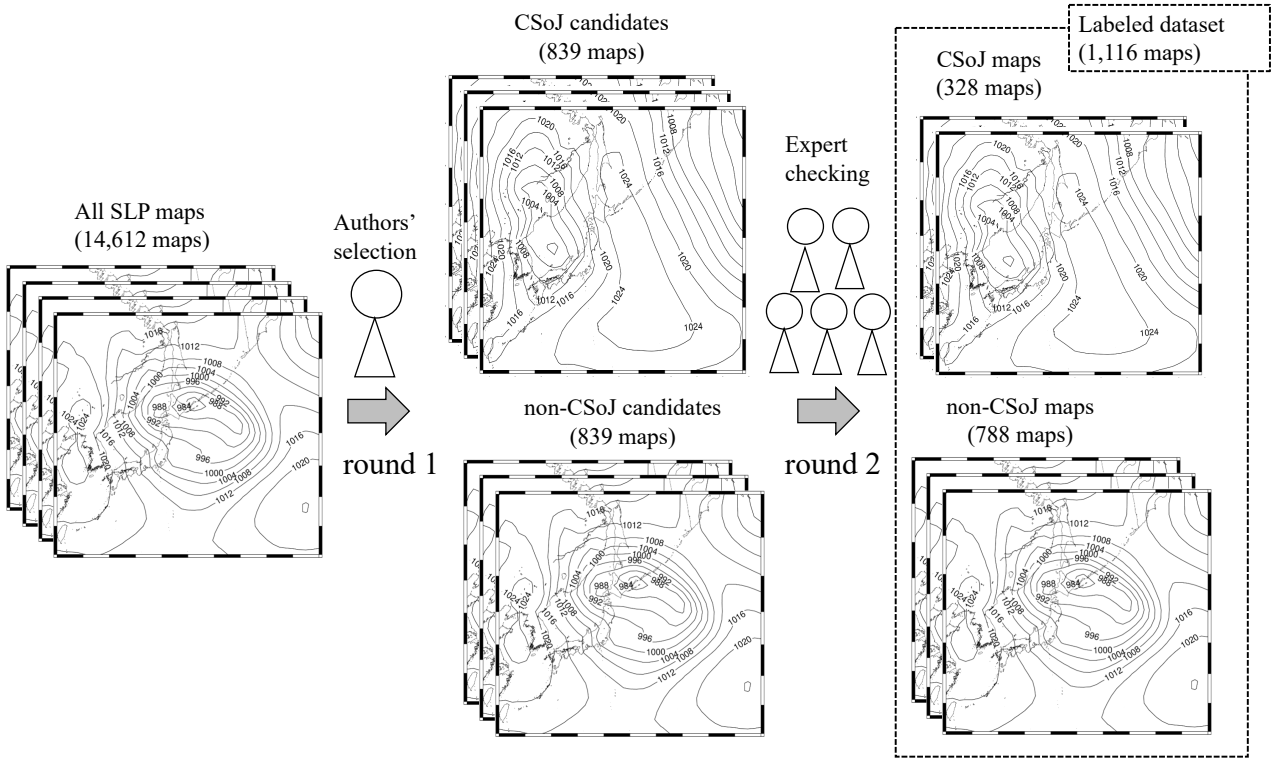
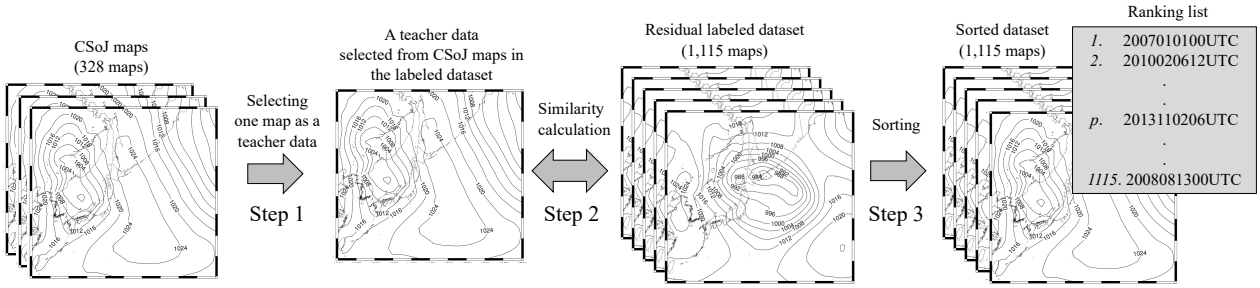


Fig. 2 Schematic of the subjective labeling process.



443

444 Fig. 3 Schematic of workflow.

445

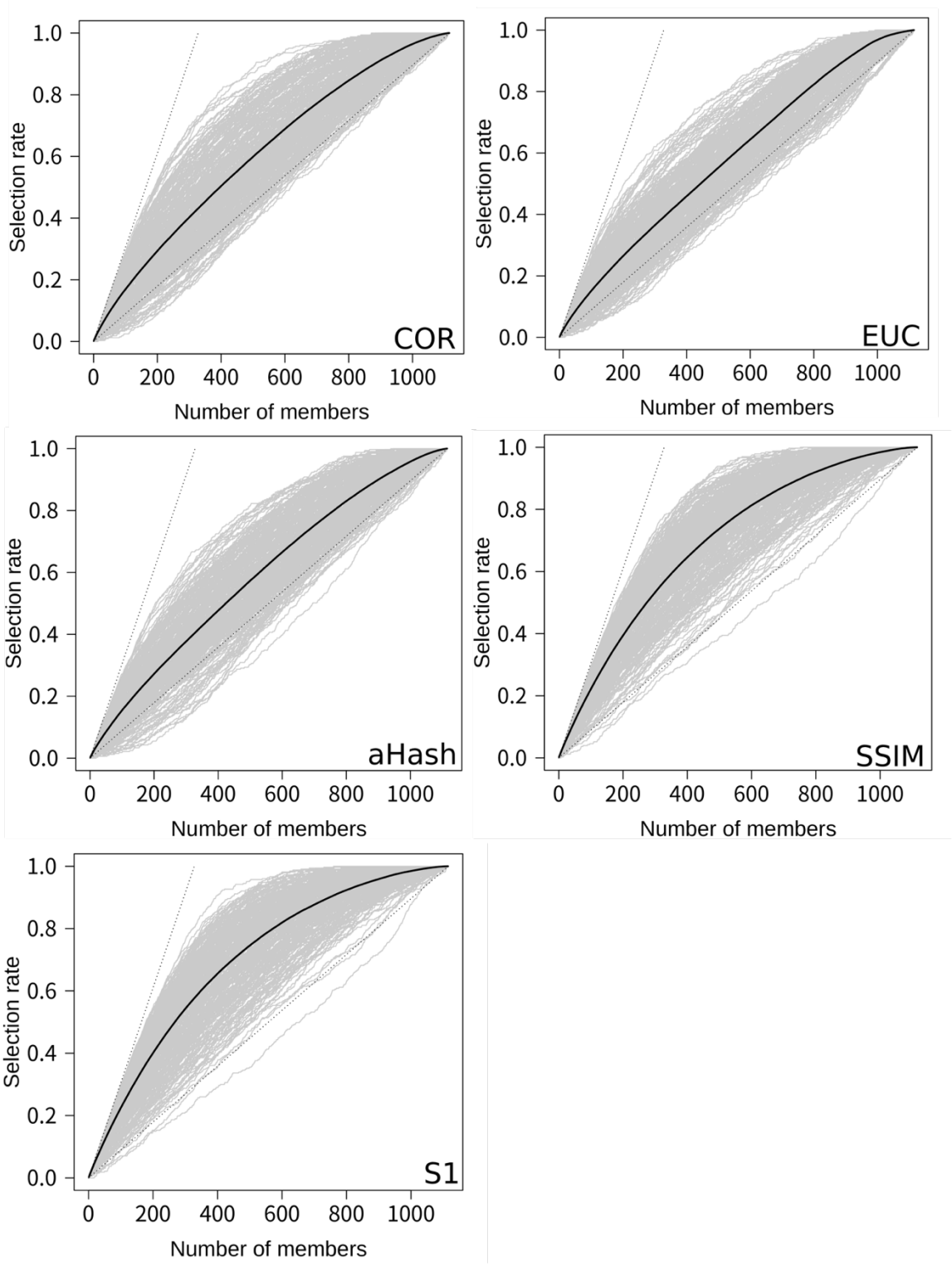


Fig. 4 Selection rate of similarity metrics. The gray line represents the curve of teacher data, the black line is the mean curve of teacher data, the steep dotted line represents the ideal transition for the selection rate (all extracted data contains the CSoJ patterns), and the gentle dotted line represents the random transition of the extraction rate.

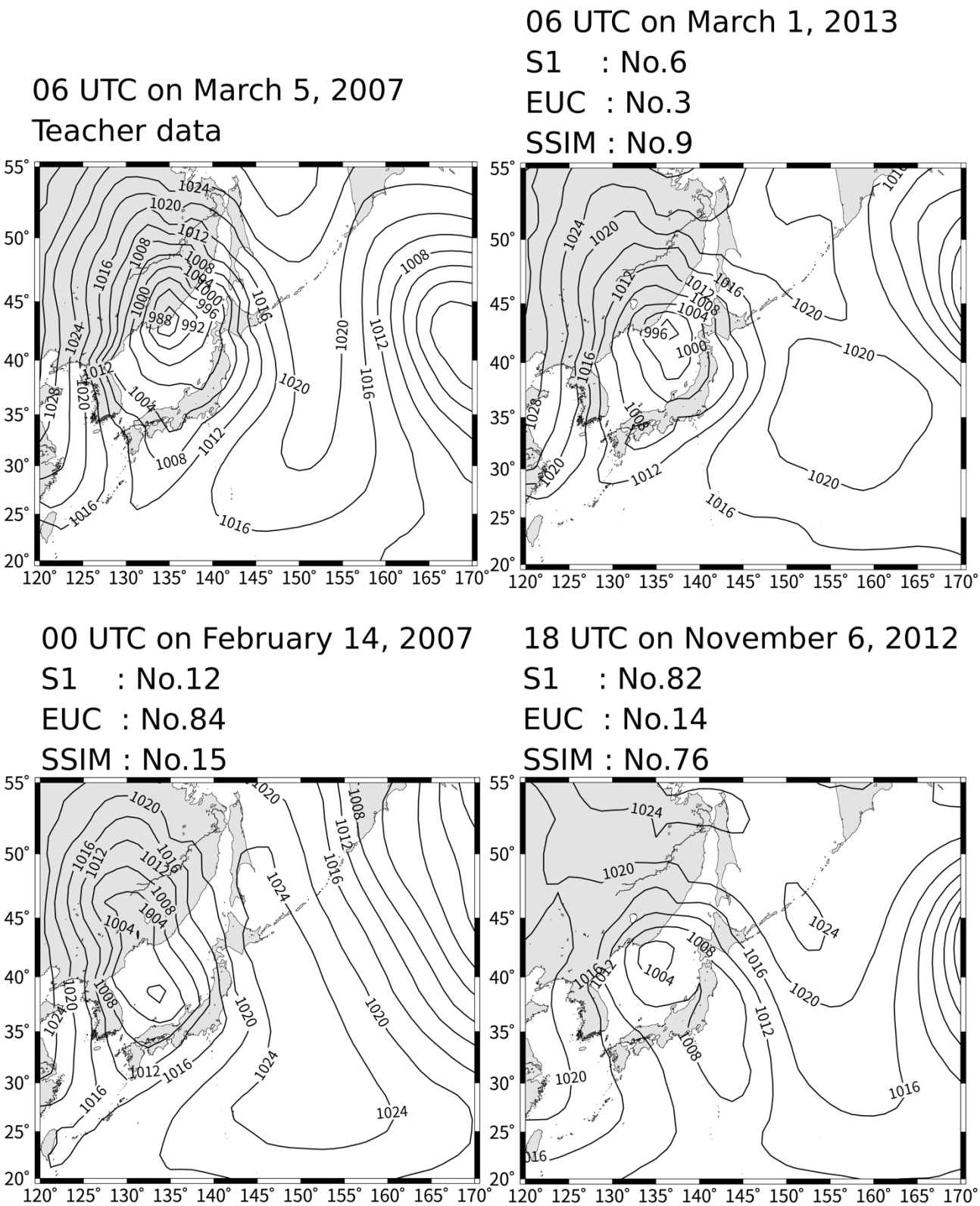


Fig. 5 Teacher and extracted SLP patterns (hPa). The numbers shown above each image denote the rank of resemblance calculated in each similarity metric. A small number of rank resemblance is associated with a high degree of similarity with the teacher data map.