Supplementary material for "Quantifying internal variability in large-ensemble climate data with the Wasserstein distance"

Yuki Yasuda[1*] and Shoichiro Kido[1]

[1]Research Institute for Earth and Information Sciences, Japan Agency for Marine-Earth Science and Technology, 3173-25 Showa-machi, Kanazawa-ku, Yokohama, 2360001, Kanagawa, Japan.

*Corresponding author(s). E-mail(s): yuki.yasuda@jamstec.go.jp;
Contributing authors: skido@jamstec.go.jp;

Computation of the information-theoretic metric γ_{info}

The computation of γ_{info} follows Sane et al. (2024). γ_{info} is obtained from the mutual information $I(X; M)$ and the Shannon entropy $H(X)$ [Eq. (7)], both of which are computed from the joint histogram of $x_{n,t}$ and m_t.

To construct the histogram, we first determine its range and the number of bins. The range is taken from the minimum to the maximum of $x_{n,t}$. Since m_t is the ensemble mean of $x_{n,t}$ [Eq. (1)], all values of m_t fall within this range. The bins are of equal width, and their number is determined by maximum a posteriori estimation (Knuth 2019). Setting the upper limit of the search to $B = 500$ bins, we compute the posterior probability at each candidate number of bins from all samples $x_{n,t}$, and then adopt the number that maximizes the posterior (Sane et al. 2024). The histogram of $x_{n,t}$ is evaluated for each candidate, and thus the total computational complexity is $O(NB)$, where $N = N_{\text{ens}} N_T$ (see Table 1).

We then construct the histogram, which approximates the joint probability distribution $p(X, M)$ of $x_{n,t}$ and m_t. The range and the number of bins determined above are used for both axes (i.e., for $x_{n,t}$ and for m_t). Denoting the bins for $x_{n,t}$ and m_t by the indices i and j, respectively, we approximate $p(X, M)$ by p_{ij}. The marginal distributions $p(X)$ and $p(M)$ are then given by $p_i := \sum_j p_{ij}$ and $p_j := \sum_i p_{ij}$, respectively.

The mutual information $I(X;M)$ and the Shannon entropy $H(X)$ are computed as follows:

$$I(X;M) = \sum_{ij} \left[p_{ij} \log_2 \left(\frac{p_{ij}}{p_i p_j} \right) \right],$$

$$H(X) = -\sum_i \left(p_i \log_2 p_i \right).$$

The metric γ_{info} follows from these quantities (Sane et al. 2024):

$$\gamma_{\text{info}} = 1 - \frac{I(X;M)}{H(X)}.$$

Generation of the synthetic climate data

The synthetic climate data (Sane et al. 2024) are independent and identically distributed in time, and the underlying distribution is Gaussian, uniform, or lognormal. The Gaussian distribution has zero mean and unit variance, the uniform distribution has support $[-1, 1]$, and the lognormal distribution has zero mean and unit variance in its logarithm. For each distribution, correlation between ensemble members was imposed by setting the Pearson correlation coefficient ρ to a fixed value in $[0, 1)$. If the time series is sufficiently long, the correlation coefficient between any pair of members is approximately ρ. We varied ρ in steps of 0.001 from 0 to 0.009, in steps of 0.01 from 0.01 to 0.99, and in steps of 0.001 from 0.99 to 0.999. We describe below how the correlated random numbers were generated.

The Gaussian random numbers were generated by setting all off-diagonal elements of the covariance matrix to ρ. The covariance matrix is $N_{\text{ens}} \times N_{\text{ens}}$, and its diagonal elements (i.e., the variances) are all unity.

The uniform random numbers were generated using a Gaussian copula (e.g., Nelsen 2006). Gaussian random numbers with a correlation coefficient of ρ were first generated as above. Applying the cumulative distribution function of the standard Gaussian, we transformed them into sequences following the standard uniform distribution. A linear transformation was then applied to change the support from $[0, 1]$ to $[-1, 1]$. Due to the nonlinearity of the cumulative distribution function, the resulting uniform random numbers have a correlation coefficient of only approximately ρ.

The lognormal random numbers were generated using the relation between the lognormal and Gaussian distributions. If two standard lognormal variables have a correlation coefficient of ρ, then the Gaussian variables obtained by the logarithmic transformation have a correlation coefficient of ρ' (e.g., Crow and Shimizu 1988):

$$\rho' = \ln\left[1 + \rho\left(e - 1\right)\right].$$

On the basis of this relation, we generated Gaussian random numbers with ρ' as above and applied the exponential transformation to them, thereby obtaining lognormal random numbers with a correlation coefficient of ρ.

Additional results for the synthetic climate data

Figures S1 and S2 show the results for the synthetic climate data based on the uniform and lognormal distributions, respectively. Note that Fig. 2 in the main text shows the results based on the Gaussian distribution. The panel layout is as in Fig. 2, and the spread among the curves indicates the sensitivity to each condition. The results depend on the shape of the probability distribution. γ_{var} is in some cases sensitive to outliers while remaining nearly independent of N_{ens} and T. The metric γ_{info} is sensitive to T, and in some cases to N_{ens} and outliers as well. Outliers affect γ_{var} except for the uniform distribution, and γ_{info} except for the lognormal. In contrast, γ_{W_1} is nearly independent of outliers and T, although it is sensitive to N_{ens} for $N_{\mathrm{ens}} \lesssim 20$. These features of γ_{W_1} are common to all three distributions.

The departure from the truth is defined as in the main text. Figure S3 shows the dependence on N_{ens} and T without outliers. Note that Fig. 3 in the main text shows the case with outliers. Without outliers (Fig. S3), γ_{var} is the least sensitive of the three metrics to N_{ens} and T, and γ_{info} is the most sensitive. As in Fig. 3, the departure of the proposed metric γ_{W_1} from the truth depends only weakly on the distribution shape, and γ_{W_1} gives stable estimates for $N_{\mathrm{ens}} \sim 40$ or more.

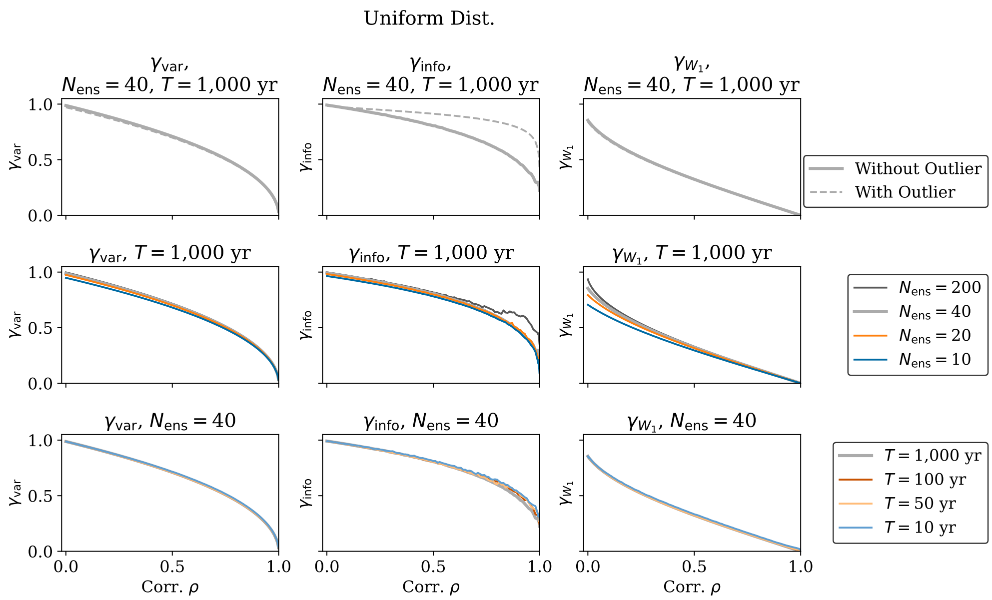

Fig. S1 Sensitivity of the three metrics for synthetic data based on the uniform distribution. Each column shows one metric (γ_{var}, γ_{info}, and γ_{W_1}, from left to right) as a function of the correlation coefficient ρ. Top: with (dashed) and without (solid) outliers, for $N_{\text{ens}} = 40$ and $T = 1{,}000$ years. Middle and bottom: four values of the ensemble size N_{ens} ($T = 1{,}000$ years) and of the data (time series) length T ($N_{\text{ens}} = 40$), respectively, both without outliers.

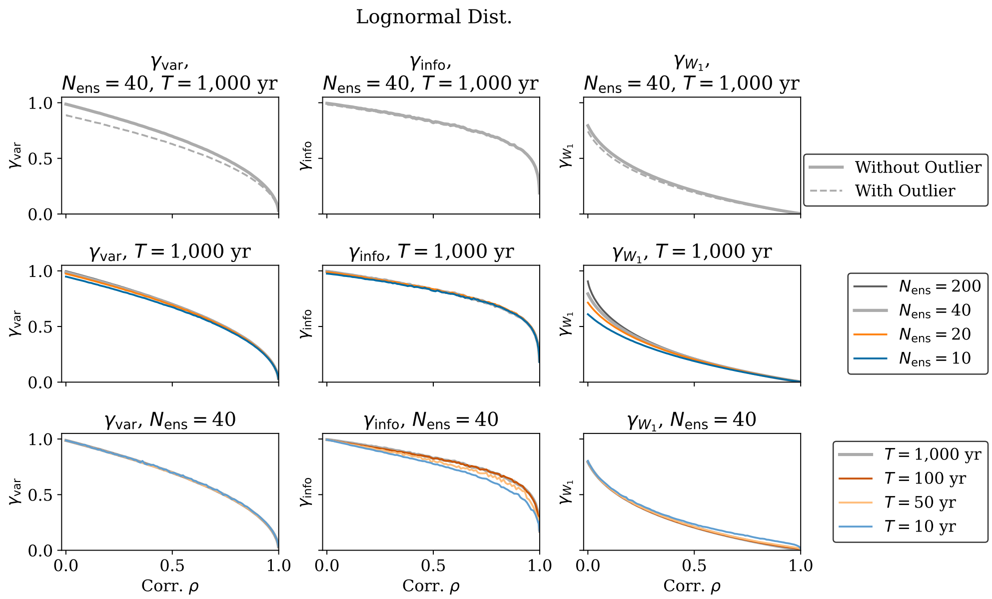

Fig. S2 As in Fig. S1, but for synthetic data based on the lognormal distribution.

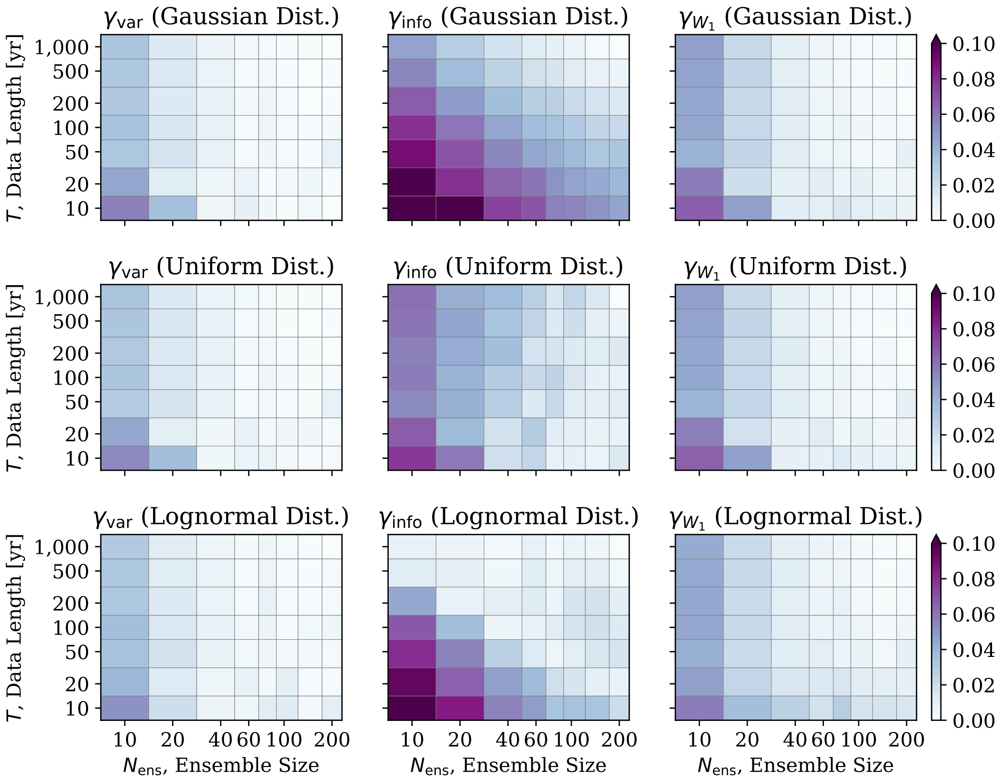

Fig. S3 Departure of the three metrics from the truth for synthetic data without outliers. Columns show the metrics ($\gamma_{\rm var}$, $\gamma_{\rm info}$, and γ_{W_1}, from left to right) and rows the Gaussian, uniform, and lognormal distributions. Each panel gives the departure on a grid of the ensemble size $N_{\rm ens}$ and the data (time series) length T, with the same color scale in all panels. The truth is the value obtained for $N_{\rm ens} = 200$ and $T = 1{,}000$ years, and the departure is the absolute difference from it integrated over $\rho \in [0, 1)$.

Additivity of the two transport costs w_{int} and w_{ext}

We restate the two transport costs and the proposed metric [Eqs. (11)–(13)]:

$$\begin{aligned}
w_{\text{int}} &= \frac{1}{N_{\text{ens}} N_T} \sum_{i=1}^{N_{\text{ens}} N_T} |x_{(i)} - m_{(i)}|, \\
w_{\text{ext}} &= \frac{1}{N_{\text{ens}} N_T} \sum_{i=1}^{N_{\text{ens}} N_T} |m_{(i)} - \overline{m}|, \\
\gamma_{W_1} &= \frac{w_{\text{int}}}{w_{\text{int}} + w_{\text{ext}}}.
\end{aligned}$$

Transporting the distribution of all samples $x_{n,t}$ directly to the total mean $\overline{m}$ gives a third transport cost, from which another metric follows:

$$\begin{aligned}
w_{\text{tot}} &= \frac{1}{N_{\text{ens}} N_T} \sum_{i=1}^{N_{\text{ens}} N_T} |x_{(i)} - \overline{m}|, \\
\widetilde{\gamma}_{W_1} &= \frac{w_{\text{int}}}{w_{\text{tot}}}.
\end{aligned}$$

The triangle inequality gives $w_{\text{tot}} \leq w_{\text{int}} + w_{\text{ext}}$. This inequality means that the cost of transporting $x_{n,t}$ directly to $\overline{m}$ is smaller than or equal to the total cost of the two-step transport, where $x_{n,t}$ is transported to m_t and then m_t to $\overline{m}$. The two metrics are therefore ordered as follows:

$$\gamma_{W_1} = \frac{w_{\text{int}}}{w_{\text{int}} + w_{\text{ext}}} \leq \frac{w_{\text{int}}}{w_{\text{tot}}} = \widetilde{\gamma}_{W_1}.$$

Equality holds in both relations if and only if $m_{(i)}$ lies between $x_{(i)}$ and $\overline{m}$ for every i, that is, if $x_{(i)} - m_{(i)}$ and $m_{(i)} - \overline{m}$ share a sign. This condition is satisfied when the quantiles of m_t are those of $x_{n,t}$ contracted toward $\overline{m}$, as in the synthetic climate data based on the uniform distribution.

In this study, we adopt γ_{W_1} rather than $\widetilde{\gamma}_{W_1}$ for two reasons. First, the denominator of γ_{W_1} separates into an internal and a forced part, as those of γ_{var} and γ_{info} do, whereas w_{tot} does not. This construction also bounds γ_{W_1} between 0 and 1. Second, γ_{W_1} never exceeds $\widetilde{\gamma}_{W_1}$, and it is thus the conservative choice for the relative contribution of internal variability.

The difference between γ_{W_1} and $\widetilde{\gamma}_{W_1}$ is small in practice. In every synthetic experiment, with and without outliers, $\widetilde{\gamma}_{W_1} - \gamma_{W_1}$ remains below 0.02 (not shown). For CESM-LE, Figs. S4 and S5 compare the two metrics. Their maps are nearly identical for both the 2 m air temperature and the total precipitation, under the 20C and the RCP8.5 forcing. The difference $\widetilde{\gamma}_{W_1} - \gamma_{W_1}$ is positive everywhere, consistent with the

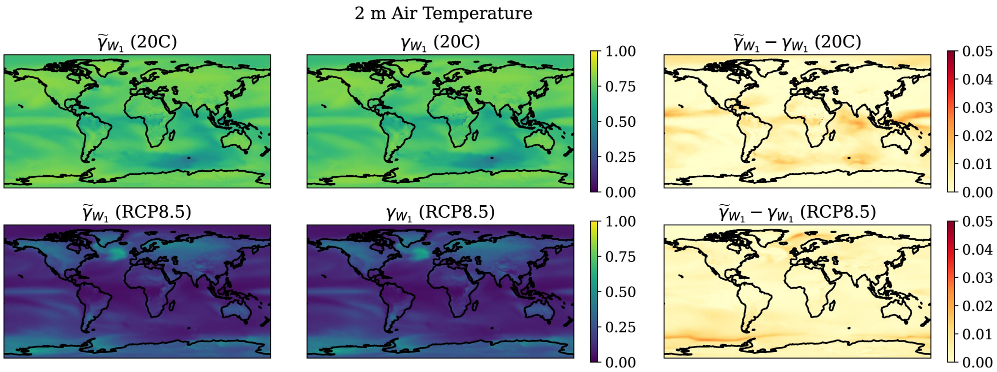

Fig. S4 Two normalizations of the proposed metric for the 2 m air temperature from CESM-LE. Columns show $\widetilde{\gamma}_{W_1}$, which is normalized by the direct transport cost w_{tot}, then γ_{W_1}, and then the difference between the two ($\widetilde{\gamma}_{W_1} - \gamma_{W_1}$). The middle column reproduces the γ_{W_1} panels of Fig. 4 in the main text for comparison. Top and bottom rows: 20th-century historical forcing (20C) and representative concentration pathway 8.5 (RCP8.5). Both metrics are computed independently at each grid point from the same 40 ensemble members.

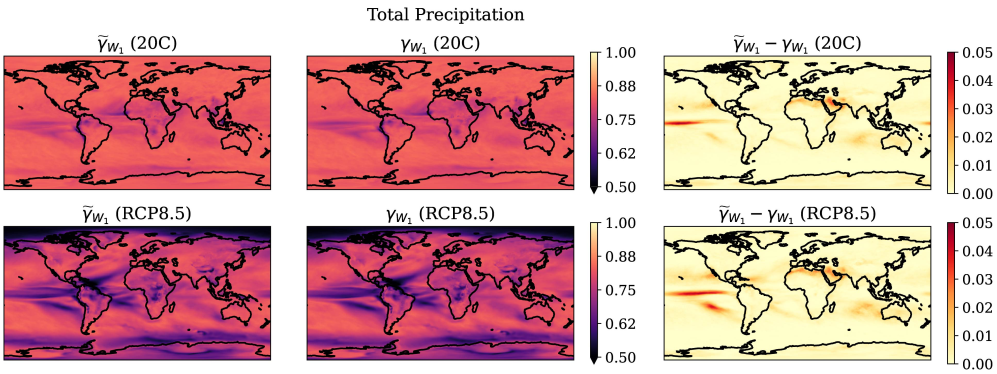

Fig. S5 As in Fig. S4, but for the total precipitation. The middle column reproduces the γ_{W_1} panels of Fig. 5 in the main text for comparison.

ordering above, and stays below 0.05. Its largest values appear in the eastern equatorial Pacific for the total precipitation, a variable whose distribution is heavy-tailed (Franzke et al. 2020). Additivity of the two transport costs is therefore not exact, but the departure from it is small enough that the two metrics γ_{W_1} and $\widetilde{\gamma}_{W_1}$ support the same conclusions.

References

Crow EL, Shimizu K. Lognormal distributions: Theory and applications. First ed. Routledge; 1988.

Franzke CLE, Barbosa S, Blender R, Fredriksen HB, Laepple T, Lambert F, et al. The Structure of Climate Variability Across Scales. Reviews of Geophysics. 2020;58(2):e2019RG000657. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019RG000657, https://doi.org/10.1029/2019RG000657.

Knuth KH. Optimal data-based binning for histograms and histogram-based probability density models. Digital Signal Processing. 2019;95:102581. https://www.sciencedirect.com/science/article/pii/S1051200419301277, https://doi.org/10.1016/j.dsp.2019.102581.

Nelsen RB. An introduction to copulas. Springer; 2006.

Sane A, Fox-Kemper B, Ullman DS. Internal Versus Forced Variability Metrics for General Circulation Models Using Information Theory. Journal of Geophysical Research: Oceans. 2024;129(5):e2023JC020101. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023JC020101, https://doi.org/10.1029/2023JC020101.