# Ranking IPCC Models Using the Wasserstein Distance

**G. Vissio[1], V. Lembo[1], V. Lucarini[1,2,3]and M. Ghil[4,5]**

[1]CEN, Meteorological Institute, University of Hamburg, Hamburg, Germany
[2]Department of Mathematics and Statistics, University of Reading, Reading, UK
[3]Centre for the Mathematics of Planet Earth, University of Reading, Reading, UK
[4]Geosciences Department and Laboratoire de Météorologie Dynamique (CNRS and IPSL), Ecole Normale Supérieure and PSL University, Paris, France
[5]Department of Atmospheric & Oceanic Sciences, University of California at Los Angeles, Los Angeles, USA

**Key Points:**

- Evaluation of climate model performance by benchmarking with reference datasets
- Climate model ranking related to the choice of variables of interest
- Highlighting model deficiencies through emphasis on climatic regions and variables

Corresponding author: Gabriele Vissio, gabriele.vissio@uni-hamburg.de

**Abstract**

We propose a methodology for evaluating the performance of climate models based on the use of the Wasserstein distance. This distance provides a rigorous way to measure quantitatively the difference between two probability distributions. The proposed approach is flexible and can be applied in any number of dimensions; it allows one to rank climate models taking into account all the moments of the distributions. Furthermore, by selecting the combination of climatic variables and the regions of interest, it is possible to highlight the deficiencies of each of the models under study. The Wasserstein distance thus enables a comprehensive evaluation of climate model skill. We apply this approach to a selected number of physical fields, ranking the models in terms of their performance in simulating them, as well as pinpointing their weaknesses in the simulation of some of the selected physical fields in specific areas of the Earth.

## 1 Introduction and motivation

Advanced climate models differ in the choice of prognostic equations and in the methods for their numerical solution, in the number of processes that are parametrized and the choice of the physical parametrizations, as well as in the way the models are initialized, to mention just their most important aspects. Comparing the performance of such models is still a major challenge for the climate modeling community (Held, 2005). Each model has its own strengths and weaknesses and, as a result, past reconstructions and future projections of climate necessarily come with model-dependent uncertainties.

Model inadequacies result from structural errors — certain processes are incorrectly represented or not represented at all — as well as from parametric uncertainties, i.e., the use of incorrect values of physical and other parameters (Lucarini, 2013; Ghil & Lucarini, 2020). Investigating the properties of multi-model ensembles is crucial for addressing climate modeling errors, while auditing climate models is essential for understanding which ones are more skillful in answering the specific climate question under study.

Testing model performance in order to advance climate modeling skill has led the community to pool its efforts within the Coupled Model Intercomparison Project (CMIP), which is currently in its sixth phase (Eyring, Bony, et al., 2016). Dozens of modeling groups have agreed by now on a concerted effort to provide numerical simulations with standardized experimental protocols representative of specified climate forcing scenarios.

The issue of best practices for model performance evaluation has naturally arisen in this setting. Such practices have concentrated essentially on either "metrics" or diagnostics. Performance metrics (Gleckler et al., 2008) have been used to rank models according to specific scalar indices that summarize overall performance, but the evaluation criterion on which such an index is based appears to be somewhat arbitrary so far.

Diagnostics, on the other hand, are process-based and designed to assess specific features of the climate system. Eyring, Righi, et al. (2016) have conducted recently an effort to bring together metrics and diagnostics in a standardized framework for climate model evaluation. Still, it seems highly desirable to have a scalar metric that summarizes the full information associated with model performance and that does satisfy the mathematical axioms associated with the concept and satisfied by the usual Euclidean distance. These axioms are listed in Text S1 of the Supplementary Information and they are satisfied by the root-mean-square distance, known as an $L_2$ metric in mathematics. The latter distance, though, is not appropriate for describing fully the difference between two distribution functions, while other metrics used in the climate sciences are not genuine distances, i.e., they do not satisfy the axioms above.

We propose a genuine metric to assess a climate model's skill by taking into account every moment of a distribution and measuring, in a much more satisfactory way, the gap between it and another distribution of reference than root-mean-square distance. The two distributions will be chosen here to describe model features, on the one hand, and the "real world," on the other, with the latter distribution being based on either raw observations or a reanalysis thereof.

Ghil (2015) originally proposed the idea of using the Wasserstein distance (hereafter WD) (Kantorovich, 2006; Dobrushin, 1970) in the context of the climate sciences as a way to generalize the traditional concept of equilibrium climate sensitivity (Ghil & Lucarini, 2020) in the presence of a time-dependent forcing, such as seasonal or anthropogenic forcing. Robin et al. (2017) used the WD to compute the difference between the snapshot attractors of the Lorenz (1984) model for different time-dependent forcings, providing a link between nonautonomous dynamical systems theory and optimal transport. Vissio and Lucarini (2018) used the WD to evaluate the skill of a stochastic parametrization for a fast-slow system. Please see Text S1 in the Supplementary Information for further background on the WD.

The WD will be calculated in a phase space defined by the physical fields we wish to take into account and it is therewith a well-suited candidate for a comprehensive and flexible way to evaluate a climate model's statistical skill. A well-known WD drawback consists in its computational requirements, which increase dramatically with the number of points needed to construct the distributions. In our methodology, following Vissio and Lucarini (2018) and Vissio (2018), these requirements are greatly reduced through data binning on a grid. Doing so one switches from the distance between distributions of points to the distance between the measures computed on the distributions themselves, which reduces the effective sample size for each distribution.

The WD-based methodology helps complement and refine the existing tests already applied in climate modeling studies, such as the space-time ranking of model performances by Flato et al. (2013), with respect to the root-mean-square-error of the median of an ensemble, with observations used as expected values, or weighting schemes like in Knutti et al. (2017). Data are presented in Sec. 2, methods in Sec. 3, results in Sec. 4, and conclusions in Sec. 5.

## 2 Data

The WD methodology is presented in Sec. 3. It is applied here to three climate fields:

- Near-surface air temperature;
- Precipitation; and
- Sea ice cover, computed from the sea ice area fraction.

The corresponding daily mean fields are available in the CMIP5 simulations for historical and RCP85 forcings (Taylor et al., 2012) and they are ranked with respect to the distance from reference daily datasets, specifically European Centre for Medium-Range Weather Forecasts Re-Analysis (ERA) Interim for the temperature (Dee et al., 2011); Global Precipitation Climatology Project (GPCP) for the precipitation (Adler et al., 2003); and Ocean and Sea Ice - Satellite Application Facility (OSI-SAF) for the sea ice cover (EUMETSAT Ocean and Sea Ice Satellite Application Facility, 2017). In order to further support the comparison and provide a benchmark, we analyzed the WD with respect to the National Center of Environmental Prediction (NCEP) Reanalysis 2 (Kanamitsu et al., 2002).

The fields are averaged on four distinct domains: (i) Global; (ii) Tropics – defined as the region between 30 S and 30 N; (iii) Northern extratropics: from 30 N to 90 N; and (iv) Arctic – used only for sea ice extent. While temperature and precipitation analyses involve a total of 30 models, taking into account sea ice extent allows to analyze just 22 models, due to available datasets. The time range spans 18 years, from January 1st, 1997 to December 31st, 2014. After the spatial averaging, the model datasets are obtained by concatenating the historical runs, from 1997 to 2005, and the RCP85 runs, from 2006 to 2014. The acronyms of the models that participated in CMIP5 and were used here appear in the figures below and are given in Table S1 of Text S2 in the Supplementary Information.

The samples used in the WD calculations are drawn by performing a Ulam (1964) discretization of the phase space involved in each separate test. To do so, a regular grid is superposed over all the datasets used in the test and its upper and lower limits, respectively, are fixed slightly above and below the maximum and minimum values among all the datasets used in it. Each dimension of the grid is then equally divided into 20 intervals; this yields $20^n$ $n$-dimensional cubes, where $n$ is the number of fields taken into account in the test. These $20^n$ hypercubes provide the sample for each test. The results we present here are weakly sensitive to the specifics of the gridding. Nonetheless, a too coarse gridding removes a lot of the information we want to retain and analysis; a too fine gridding, instead, increases substantially the computing requirements, without making much statistical sense.

In order to highlight the flexibility and reliability of the method, we are going to calculate the WD distances in one-, two- and three-dimensional phase space, and work with different field combinations averaged over distinct areas of the Earth.

## 3 Wasserstein distance (WD)

Our objective is to create a ranking of the CMIP5 IPCC models based on their skill to reproduce the statistical properties of selected physical quantities. The reference distribution for these quantities is given by reanalysis and observational datasets, as explained in Sec. 2; their WD to these datasets (Kantorovich, 2006; Villani, 2009) is a measure of the models' ability to reproduce these reference distributions. One can also describe this

distance as the minimum "effort" to morph one distribution into the other (Monge, 1781). We present below a very simplified account of the theory.

The optimal transport cost (Villani, 2009) is defined as the minimum cost to move the set of points from one distribution to another into an n-dimensional phase space. In the case of two discrete distributions, we write their measures $\mu$ and $\nu$ as

$$\mu = \sum_{i=1}^{n} \mu_i \delta_{x_i}, \qquad \nu = \sum_{i=1}^{n} \nu_i \delta_{y_i}; \tag{1}$$

here $\delta_{x_i}$ and $\delta_{y_i}$ are Dirac measures associated with a pair of points $(x_i, y_i)$, whose fractional mass is $(\mu_i, \nu_i)$, respectively, and $\sum_{i=1}^{n} \mu_i = \sum_{j=1}^{n} \nu_j = 1$. $n$ is the number of dimensions in the phase space in which we compute the WD. Using the definition of Euclidean distance

$$d(\mu, \nu) = \left[ \sum_{i=1}^{n} (x_i - y_i)^2 \right]^{\frac{1}{2}}, \tag{2}$$

we can write down the quadratic WD for discrete distributions:

$$W_2(\mu, \nu) = \left\{ \inf_{\gamma_{ij}} \sum_{i,j} \gamma_{ij} [d(x_i, y_j)]^2 \right\}^{\frac{1}{2}}. \tag{3}$$

where $\gamma_{ij}$ is the fraction of mass transported from $x_i$ to $y_j$ and $d(x_i, y_j)$ is the Euclidean distance between a single pair of locations.

We perform the Ulam discretization described in Sec. 2 — i.e. data binning on a grid chosen to have a resolution of 20 intervals per side, as mentioned above — that allows us shift from the distance between different distributions of points to the distance between the measures related to those distributions. We thus proceed to quantify to what extent the measure of the observations and reanalysis from Sec. 2, projected on the variables of interest, differs from the corresponding measures for the climate models.

The estimate of the coarse-grained probability of being in a specific grid box is given by the time fraction spent in that box (Ott, 1993; Strogatz, 2015). In fact, the WD does provide robust results even with a very coarse grid (Vissio & Lucarini, 2018; Vissio, 2018). Therefore, in the case at hand, the locations $x_i$ and $y_j$ will indicate the cubes' centroids, while $\gamma_{ij}$ indicate the corresponding densities of points. To further simplify the computations, we exclude all the grid boxes containing no points at all. Finally, we "renormalize" the densities, dividing the value obtained by the number of grid intervals per side; therefore, the one-, two- and three-dimensional WDs take values between a minimum of 0 and a maximum equal to 1, $\sqrt{2}$ and $\sqrt{3}$, respectively.

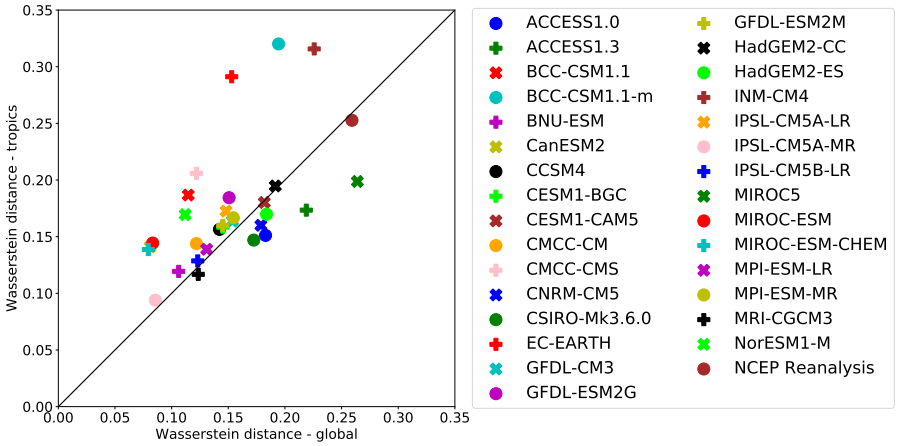


**Figure 1.** Two-dimensional Wasserstein distance (WD) for the temperature and precipitation fields, averaged over the globe (horizontal axis) and over the Tropics (vertical axis). The acronyms of the models used are spelled out in Text S2 of the Supplementary Information.

We used a suitably modified version of the Matlab software written by G. Peyré — available at `http://www.numerical-tours.com/matlab/optimaltransp_1_linprog/` — to perform the calculations. The modifications include the data binning and the estimation of the measures, as well as adapting to a dimension $n \geq 2$.

## 4 Ranking the models

Figure 1 shows the WD calculated in the two-dimensional phase space composed by the temperature and precipitation fields, averaged over the whole Earth and the Tropics, for each CMIP5 model. In order to provide a benchmark, we chose to include the WD results between the NCEP reanalysis and the references of ERA and GPCP presented in Sec. 2 for the two fields, respectively.

Somewhat surprisingly, the NCEP reanalysis yields the largest values in both distances. Thus, the average CMIP5 distance to the ERA $\otimes$ GPCP reference is 0.149, while the NCEP distance is 0.259, exceeded only by the value 0.264 given by the MIROC5 model; see Table S1 in the Supplementary Information for the list of models. Note that the one-dimensional WDs of the NCEP Reanalysis for the globally averaged temperature and precipitation equal 0.033 and 0.255, respectively. Given the well-known difficulties with simulating the very rough precipitation field by using the still fairly coarse CMIP5 mod-

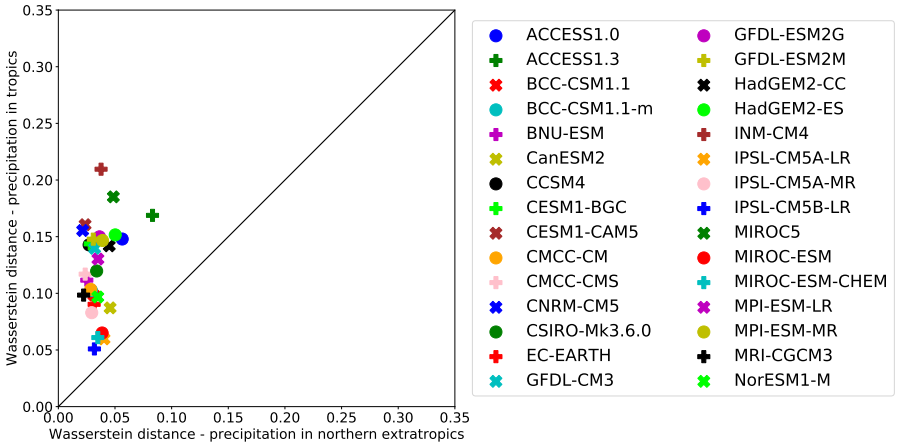


**Figure 2.** One-dimensional WD for precipitation averaged over the Northern extratropics (from 30 N to 90 N) on the horizontal axis and over the Tropics (from 30 S to 30 N on the vertical axis).

els (Neelin et al., 2013; Mehran et al., 2014), it is natural to assume that both the ERA and NCEP reanalyses are mostly inadequate in representing the statistics of precipitation. The great discrepancy in WD between the distribution of reference and the NCEP Reanalysis points to the overall accuracy reached by CMIP5 simulations when dealing with global averages of temperatures and precipitation.

We evaluate next the problems still encountered by CMIP5 models in reproducing key aspects of tropical dynamics (Tian & Dong, 2020). Averaging the data over the Tropics, we obtain the ranking on the vertical axis in Fig. 1. The large values of the WD distances equal on average 0.173, excluding the NCEP Reanalysis, and underline the poorer CMIP5 model performances in this region. With few exceptions, the models seem less reliable in the Tropics, where three of the models do exceed the NCEP Reanalysis distance.

Focusing on the relative performance of temperature and precipitation in the Tropics vs the Northern Hemisphere extratropics (30 N–90 N), Figs. 2 and 3 illustrate one-dimensional WDs computed in the former vs the latter region. Using the diagonal line indicating equal values for the two distances as a reference, we can easily check in Fig. 2 that, for all CMIP5 models, the precipitation field is less well reproduced in the Tropics than in the extratropics: it is well known that it is extremely challenging to repro-

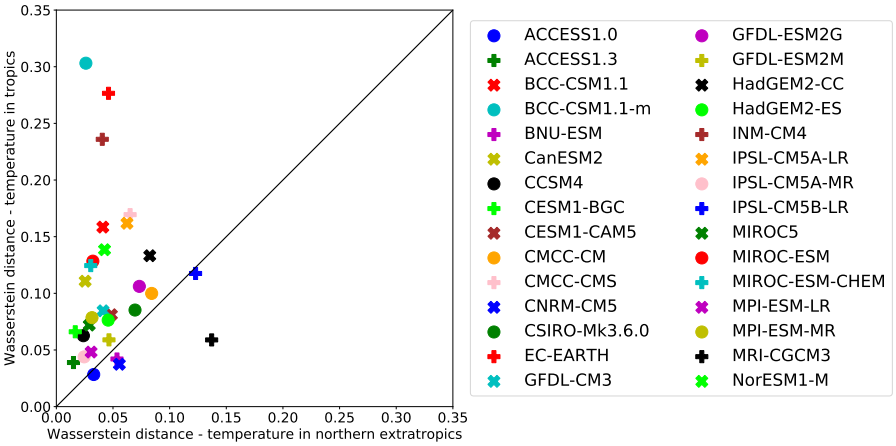


**Figure 3.** Same as Fig. 2 but for the temperature field.

duce accurately the statistics of by-and-large convection-driven precipitation, since the choice of the parametrization schemes and their tuning plays an essential role. The situation for the temperature field is similar but less uniformly so: while in Fig. 2 all the results cluster above the diagonal but roughly below WD $\simeq$ 0.2, the scatter in Fig. 3 is larger, with some results below the diagonal and some between 0.2 $\lessapprox$ WD $\lessapprox$ 0.3.

Figure 4 shows the scatter diagram of one-dimensional WDs for the precipitation in the Tropics vs the WDs of sea ice extent in the Arctic. Arctic sea ice cover is a very important indicator of the state of both hydrosphere and cryosphere, as well as of their mutual coupling; it is overestimated in CMIP5 models during the winter and spring seasons (Randall et al., 2007; Flato et al., 2013).

Figure 4 demonstrates that the sea ice cover in the models is closer to the observations than the tropical precipitation in 12 CMIP5 models out of the 22 examined. Nevertheless, 7 models better describe tropical precipitation than sea ice extent in the Arctic, while 3 models have a similar — and relatively low — WD for both fields. This test indicates that a correct representation of the statistics of these two fields is still quite challenging across the spectrum of climate models at the present time.

We compare next the performance of the CMIP5 models with respect to three different rankings. First, the three-dimensional WD is computed taking into account three physical quantities: globally averaged temperature and precipitation, along with sea ice

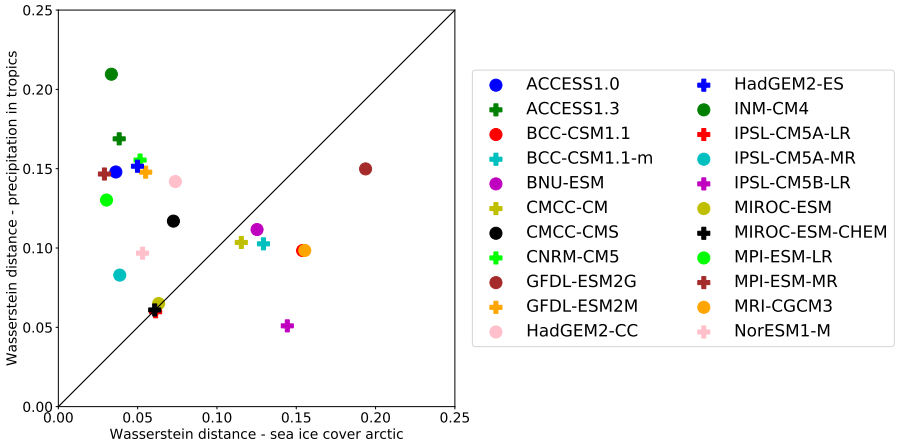


**Figure 4.** One-dimensional WDs of average precipitation in the Tropics vs the average sea ice extent in the Arctic.

extent in the Arctic. Note that, to ease the interpretation of Fig. 5, the models are listed on the vertical axis according to the rank provided by this methodology.

The model ranking introduced herein is further compared with the rankings based on the first two moments of the distribution of reference. For each of the three physical quantities above, we compute the normalized mean, taking the absolute value of the difference between the mean of the distribution of the model field and that of the reference field, and dividing this difference by the standard deviation of the distribution of reference. The three means for the three fields are then averaged and the same procedure is repeated for the normalized standard deviation.

We can see that the models' performance is quite different depending on the ranking being used. As an example, we focus on the BCC-CSM1.1 and BCC-CSM1.1-m models. The ranking based on the mean shows a rather good performance for both, with positions 7 and 10, respectively; nevertheless, they occupy positions 16 and 21 in the WD ranking. The latter low positions are due to their bad performances when it comes to standard deviation, where the two come last.

The reverse instance is also clear by looking at those models that, while performing well in terms of variability, occupy lower rankings based on the WD due to their poor performance in the mean; see, for instance, the case of MPI-ESM-MR, with position 1

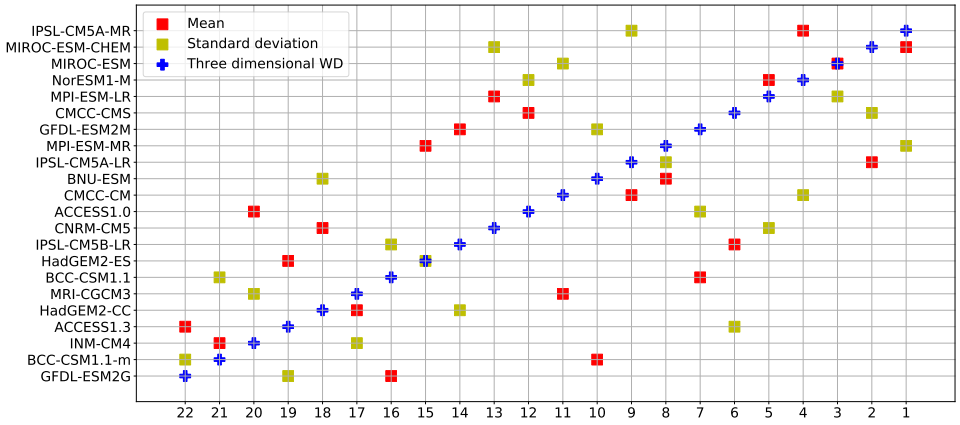


**Figure 5.** Comparing 22 CMIP5 models (vertical axis) vs their positions in the ranking (horizontal axis): (a) three-dimensional WD – heavy blue '+' sign; (b) mean WD – red filled square; and (c) standard deviation of WDs – yellow filled square. See text for explanations. See Tables S2-S4 in Supplementary Information for detailed results.

in the standard deviation, 8 in WD, and 15 in the mean. The WD score accounts for the information carried by the whole distribution — i.e., by the mean, standard deviation and higher moments — and clearly balances out the first and second moment thereof.

A more peculiar instance is provided by HadGEM2-CC and HadGEM2-ES, which rank in this order for both the mean (17th and 19th) and the standard deviation (14th and 15th), but in the reverse order in the WD ranking (18th and 15th). This apparent paradox could be due to the presence of nontrivial second-order correlations between the variables or from the effect of higher moments of the distributions.

Note that, for the 18-year time interval studied herein (1997–2014), the results obtained applying the WD approach in three-dimensional phase space are not very different from those given by averaging the three corresponding one-dimensional distances. This agreement is due to the unimodality of the distributions taken into account and things would be different, for instance, if one were studying a paleoclimate setting that includes bimodality of the sea ice cover but not of the temperature field. In any case, the full application of the multi-dimensional WD leads to more robust results, as all correlations between the variables are taken into consideration.

## 5 Conclusions

We have proposed a new methodology to study the performance of climate models based on the computation of the Wasserstein distance (WD) between the multidimensional distributions of suitably chosen climatic fields of reference datasets and those of the models of interest. This method takes into account all the moments of the distributions and it is, therefore, more informative and more robust than ranking methods based on means or variances alone. The methodology is flexible as it allows one to consider several variables at the same time; it thus has the potential of disentangling the effect of the correlation between different climatic quantities.

The proposed methodology has been proven to be effective in pointing to climate modeling problems related to the representation of quantities like precipitation or sea ice extent over limited areas, such as the Tropics and the Arctic, respectively; see again Figs. 2 and 3. Furthermore, this methodology can be applied to studying model performance for a given climatic variable over different spatial domains, as seen in Figs. 1–4, as well as relative model performance for different fields, as seen in Fig. 4. This flexibility can help guide attempts at model improvements by providing robust diagnostics of the least well simulated field — temperature, precipitation or sea ice extent — or region, namely either hemisphere, the Tropics or the Arctic.

Such a method, taking into account the whole distribution of the statistics and not just one representative number, like its mean or standard deviation, is complementary to those already in use, allowing for a deeper understanding of the models' performance and the reasons behind their inadequacies. Unlike most evaluation methods for climate models used so far (Flato et al., 2013), this approach does not rely on correlations, variances or mean square errors, and thus it does not focus only on standard measures of variability; rather, it shows quantitatively if a model does a good job in reproducing the desired statistics — including every moment of the distributions — and, more importantly, it allows one to compare several different fields at the same time, checking quantitatively differences in the aforementioned statistics among different models and fields.

Throughout the paper, we have shown the application of this approach to different physical fields, providing a ranking of CMIP5 models for specific sets of fields, as well as a way to highlight model weaknesses to help focus the honing of climate models. Get-

ting more reliable models will lead to better simulations and, therefore, to more accurate climate predictions.

**Acknowledgments**

It is a pleasure to acknowledge our data sources: GPCP and NCEP Reanalysis 2 data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at `https://www.esrl.noaa.gov/psd/` and OSI-SAF data provided by the EUMETSAT Ocean and Sea Ice Satellite Application Facility - Global sea ice concentration climate data record 1979-2015 (v2.0, 2017). The present paper is TiPES contribution #30; the TiPES (Tipping Points in the Earth System) project has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 820970. Work on this paper has also been supported by the EIT Climate-KIC; EIT Climate-KIC is supported by the European Institute of Innovation & Technology (EIT), a body of the European Union, under Grant Agreement No. 190733.

## References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P., Janowiak, J., ... Arkin, P. (2003). The Version 2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology*, *4*, 1147–1167.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., ... Vitart, F. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553-597.

Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, *15*(3), 458–486.

EUMETSAT Ocean and Sea Ice Satellite Application Facility. (2017). *Global sea ice concentration climate data record 1979-2015 (v2.0, 2017). Norwegian and Danish Meteorological Institutes.* `http://osisaf.met.no.` doi: 10.15770/EUM_SAF_OSI_0008

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model*

*Development*, *9*, 10539–10583.

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., ... Williams, K. D. (2016). ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *9*, 1747–1802.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., ... Rummukainen, M. (2013). Evaluation of climate models. In T. Stocker et al. (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge, UK and NY, USA: Cambridge University Press.

Ghil, M. (2015). A mathematical theory of climate sensitivity or, How to deal with both anthropogenic forcing and natural variability? In C.-P. Chang, M. Ghil, M. Latif, & J. Wallace (Eds.), *Climate Change: Multidecadal and Beyond* (Vol. 6, p. 31-52). Singapore: World Scientific Publishing Co.

Ghil, M., & Lucarini, V. (2020). The physics of climate variability and climate change. *Reviews of Modern Physics*, *in press*, arXiv:1910.00583.

Gleckler, P. J., Taylor, K. E., & Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research*, *113*, D06104.

Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, *86*, 1609–1614.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M., & Potter, G. L. (2002). NCEP-DOE AMIP-II Reanalysis (R-2). *Bulletin of the American Meteorological Society*, *Nov 2002*, 1631-1643.

Kantorovich, L. V. (2006). On the translocation of masses. *Journal of Mathematical Sciences*, *133*(4), 1381–1382. (originally published in Doklady Akademii Nauk SSSR, 37 (7–8), 199–201 (1942).)

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*, 1909–1918.

Lorenz, E. N. (1984). Irregularity: A fundamental property of the atmosphere. *Tellus A*, *36*(2), 98–110.

Lucarini, V. (2013). Modeling complexity: the case of climate science. In U. Gähde,

S. Hartmann, & J. Wolf (Eds.), *Models, simulations, and the reduction of complexity* (p. 229-254). De Gruyter.

Mehran, A., AghaKouchak, A., & Phillips, T. J. (2014). Evaluation of cmip5 continental precipitation simulations relative to satellite-based gauge-adjusted observations. *Journal of Geophysical Research: Atmospheres*, *119*(4), 1695-–1707.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, 666-704.

Neelin, J. D., Langenbrunner, B., Meyerson, J. E., Hall, A., & Berg, N. (2013). California winter precipitation change under global warming in the coupled model intercomparison project phase 5 ensemble. *Journal of Climate*, *26*(17), 6238–6256.

Ott, E. (1993). *Chaos in Dynamical Systems*. Cambridge, UK: Cambridge University Press.

Randall, D., Wood, R., Bony, S., Colman, R., Fichefet, T., Fyfe, J., ... Taylor, K. (2007). Climate Models and Their Evaluation. In S. Solomon et al. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and NY, USA: Cambridge University Press.

Robin, Y., Yiou, P., & Naveau, P. (2017). Detecting changes in forced climate attractors with Wasserstein distance. *Nonlinear Processes in Geophysics*, *24*, 393-405.

Strogatz, S. H. (2015). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering, 2nd Edition.* Boulder, CO: Westview Press.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 485–498.

Tian, B., & Dong, X. (2020). The Double-ITCZ Bias in CMIP3, CMIP5, and CMIP6 Models Based on Annual Mean Precipitation. *Geophysical Research Letters*, *47*.

Ulam, S. M. (1964). *Problems in Modern Mathematics*. New York: Science Edition Wiley.

Villani, C. (2009). *Optimal Transport: Old and New.* Berlin Heidelberg, Germany: Springer-Verlag.

Vissio, G. (2018). Statistical mechanical methods for parametrization in geophysical fluid dynamics. *Reports on Earth System Science*, *212*.

Vissio, G., & Lucarini, V. (2018). Evaluating a stochastic parametrization for a fast–slow system using the Wasserstein distance. *Nonlinear Processes in Geophysics*, *25*, 413-427.