Using Artificial Intelligence to identify CMIP6 models from daily SLP maps

Pascal Yiou, Soulivanh Thao

Using Artificial Intelligence to identify CMIP6 models from daily SLP maps

Pascal Yiou[1*] and Soulivanh Thao[1]

[1]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212 CEA-CNRS-UVSQ, IPSL and U. Paris-Saclay, CE l'Orme des Merisiers, Gif-sur-Yvette, 91191, France.

*Corresponding author(s). E-mail(s): pascal.yiou@lsce.ipsl.fr;
Contributing authors: soulivanh.thao@lsce.ipsl.fr;

Abstract

Large databases of climate model simulations are essential to sample climate variability and estimate how it can evolve in any future. The chaotic nature of climate has motivated the simulation of large ensembles of simulations, which sample the uncertainty due to internal climate variability in single models. Exploiting large ensembles (for impact or attribution studies) implicitly relies on the hypothesis that simulations are interchangeable. This is not the case for variables like temperature, due to biases (which can be corrected). Some synoptic fields, like SLP, do not yield obvious biases, which might justify their use to enrich reanalysis data. In this paper, we examine this hypothesis through a neural network classification approach. The goal is to determine whether it is possible to recognize a climate model (among 16 models and a reanalysis) from one single sea-level pressure (SLP) map over the North Atlantic. We find that models are highly identifiable in the summer (and less in other seasons), while SLP average structures are very similar. From this classification, we identify sororities of climate models, and investigate how climate change can affect SLP daily patterns toward the end of the 21st century. This study allows identifying which climate models could be used as input for artificial intelligence model forecasts.

Keywords: Neural Network, classification, Climate models, SLP

1 Introduction

Describing and understanding climate variability poses major computational, statistical and physical challenges. Some of the issues are treated by considering ensembles of coupled global climate model (GCM) simulations, i.e., with several climate models, and running several simulations. Then averages and other statistical moments are computed from those large datasets [1]. An other ambition that motivates the generation and use of those large climate datasets is to consider them as surrogates of the real world to learn how to forecast the weather [2], or provide extreme event attribution statements [3]. Those endeavors (computing statistical moments, transfer learning) make the implicit assumption that all climate models yield similar probability distributions, and that those statistical properties are the same as observations. There are many ways of removing model biases, in order to get close to this important assumption [4, 5], but those methods are rarely devised for fields (with space and time).

Even if a meteorological field obtained from a GCM does not yield obvious biases, the question we want to address is whether it is possible recognize a GCM from *one* daily map, like sea-level pressure (SLP). If one can provide such a detection, then learning from one GCM is certainly useless for other models. If GCMs cannot be distinguished, then one has a good case for pooling them in order to enlarge the set of observed climate variability that is limited to a few decades. One of the motivations of this paper stems from the idea of transfer learning [6, 7], from ensembles of climate model simulations, to actual weather predictions. In this perspective, learning from a daily timescale is crucial, in order to investigate extreme events [8].

In this paper, we consider the whole CMIP6 archive [9], which contains 47 models that provide data with daily time steps (Figure A1). We consider daily SLP over the North Atlantic, between 1970 and 2000. We determine whether it is possible to recognize a model name from *one* daily SLP map (over the North Atlantic). We use a simple artificial intelligence (AI) algorithm of classification (a Multi-Layer-Perceptron neural network [10]) without any particular tuning to learn how to recognize a model from SLP features, for each season (summer, autumn, winter and spring). The classification is done along with the ERA5 reanalysis [11]. Hence, this study extends the garment identification AI challenge [12] to a more sophisticated setting. The analysis protocol is described in the methods section.

If an AI algorithm cannot tell models apart (from daily SLP fields) or from reanalyses, then pooling model SLP can be considered a promising in order to enlarge sample size, in order to build statistical confidence. If an AI algorithm *can* tell models apart, even though their sample means and standard deviations are similar, then it is not possible to learn, even from a large simulation ensemble of one model, to infer anything on the real world.

We evaluate whether GCMs can be identified from simulations that are provided in a different setting from the learning set (different ocean model, different atmospheric parameterizations, etc.). This study revisits the analyses of [13, 14] by focusing on daily time scales, which adds a major difficulty to the classification problem, due to meteorological variability.

We determine whether a warmer climate according to SSP scenarios [15] affects the weather patterns of climate models. This is done with the SSP5-8.5 scenario for end-of-the century (2070–2100) period. The Methods section explains the analysis protocols.

2 Results

2.1 SLP Classification of 16 models

Each daily SLP map was classified onto 16 GCMs and the ERA5 [11] reanalysis (see Methods section, Table A1). The classification was done for the four seasons (Summer: June-July-August (JJA); Fall: September-October-November (SON); Winter: December-January-February (DJF); Spring: March-April-May (MAM)), due to the seasonality of the atmospheric circulation [16].

The training set used ≈ 22 years for each model, and the validation set used the remaining ≈ 8 years. The training was repeated 20 times, so that 20 AI models are obtained. We report the resulting scores on the validation periods in Figure 1. This reflects the variability of classification scores due to the algorithm itself. In the sequel, we will keep the AI models that yield the best overall scores.

The summer (June-July-August) classification of 16 models and ERA5 reanalysis shows that the AI algorithm can recognize seven models with a success probability larger than 0.6 (Figure 1a). It is possible to enhance the classification score by by repeating the neural network training with random samples of 9 GCMs rather than 16. This guarantees that the classification scores exceed 80% for daily SLP (not shown). This procedure is not discussed here, as this study is based on the minimal achievable result with AI and avoiding technical tuning procedures.

For the other seasons (spring, autumn and winter) the AI algorithm classification scores are lower, and rarely pass the 0.6 value (Figure 1b–d). This means that the SLP from 17 models (including ERA5) are difficult to distinguish from one model to another. For DJF, four GCMs can be classified with probabilities larger than 0.6, although the ERA5 reanalysis is poorly classified. For intermediate seasons, three models (for SON) or two models (for MAM) can be classified by probabilities exceeding 0.6. The SLP from the NorCPM1 model is consistently well classified across all seasons. Conversely, the EC-Earth3 model yields consistent low classification probability rates across seasons. The ERA5 reanalysis does not stand out as "recognizable" among the CMIP6 models.

A test procedure with a distinct run for each model and the best AI model in Figure 1 is performed. Figure 2 shows the classification probability distributions onto the 16 CMIP6 models (Table A1) and ERA5, with different historical runs (the test sets). The diagonals indicate whether AI model clearly identifies models from daily SLP fields. We note that, with the best AI model (from Figure 1), the JJA true positive rates exceed $p = 0.6$, which means that it is possible to identify a GCM from a daily SLP map, more that 60% of the time, which is better than tossing a coin (and higher than $p_0 = 1/17$ if all models/ERA5 are equiprobable). We verified that summer SLP can be recognized, for any run of the training models (not shown) with probabilities exceeding 0.6. This means that internal variability does not alter the SLP

identification process. We also verified that no obvious bias among GCMs explained such a classifiability by comparing mean SLP and standard deviations (SI Figure 2). The spatial structures of SLP (and its variability) are very similar between GCMs, and naked eyes would have difficulties identifying a model from a single daily SLP map.

Those classification probabilities drop during intermediate seasons (SON and MAM), although models are can still be identified, albeit with scores lower than $p = 0.3$. GCMs (or ERA5) cannot be identified from winter (DJF) daily SLP maps (Figure 2c). The ERA5 reanalysis is not more identifiable than the GCMs. Therefore, North Atlantic SLP maps of GCMs can hardly be distinguished from reanalyses, especially in the winter.

For all seasons, the EC-Earth model is often confused with the ERA5 reanalysis (Figure 2), which is explained by the fact that they are based on the same atmospheric model [17].

2.2 Identification of GCM sororities

Starting from a classification of 16 models in Table A1, we classify the "sister" models identified in Figure A1 (black bars). Those GCMs are produced by the same research groups, but can yield different horizontal resolutions (shown in Supplementary Table 1), or contain different physical configurations [1]. The goal is to check whether the classification performed on the reference models work for "sister" models, especially for the summer season.

With a couple of exceptions, sororities can be identified with SLP classification (Figure 3). When GCMs from different research institutes (e.g. UKESM1 and HadGEM3 models) share the same atmospheric model (e.g., MOHC) and yield similar horizontal resolutions, then they tend to be classified to HadGem3-GC31-LL, especially in JJA. Classifiability does not increase during other seasons. Similarly, models of the EC-Earth consortium (EC-Earth3-Veg and EC-Earth3-Veg-LR) simulations are classified to the EC-Earth3 model. The CNRM-ESM2-1 and reference CNRS-ESM1 GCMs differ from their atmospheric chemistry models. This difference does not affect their SLP sorority in the summer. The NorESM2-LM and NorCPM1 GCMs essentially differ from their vertical resolution, but their sorority can be identified.

The horizontal resolution of MPI-ESM1-2-HR is almost twice the resolution of MPI-ESM1-2-LR (all submodels are the same, though). This explains that those two GCMs cannot be classified onto one another in the summer, because the high-resolution model (which has a fairly high resemblance to ERA5) can yield SLP patterns that are not obtained with the low reference GCM. Similarly, the FGOALS-I3 model is nearly twice the resolution of FGOALS-g3, with a different atmospheric code. This also explains why the summer SLP of the two models (from the same institution) do not seem similar.

Therefore, the daily SLP maps in the summer season (JJA) allow identifying GCM sororities (Figure 3a) when horizontal resolutions are comparable, while this is much less clear for other seasons. This implies that the atmospheric circulation from models of the same family can be "pooled" in order to increase ensemble sizes. This feature is linked to the intrinsic resolution of the atmospheric model, because all models and ERA5 were re-interpolated on a $1° \times 1°$ degree grid.

2.3 Influence of climate change

We investigate the SLP classification from the 16 models in Table A1 toward the end of the 21st century in scenario simulations [9, 15]. The scenario we consider here is SSP5-8.5, and we extract the SLP from 2070 to 2100.

The probabilities along the diagonal in JJA (Figure 4a) slightly decrease in the summer, implying a slight increase of "misclassification". This suggests that some models witness the appearance of new SLP patterns (with respect to their own 1970–2000 behavior) towards the end of the 21st century, and that those patterns were sampled in other GCMs.

The highest classification rates for intermediate seasons (SON and MAM) still appear on the diagonals of Figure 4bd (the NorCPM1 model does not propose daily SLP for SSP simulations and hence does not appear). This identification no longer holds for winter SLP (Figure 4c). The overall similarity between Figures 2 and 4 suggests that the SLP patterns of climate models are barely affected by climate change, even in an extreme SSP5-8.5 scenario. Other SSPs show a similar behavior (not shown).

There is one exception to this stability: the NESM3 model whose SSP5-8.5 winter classification never finds SLP patterns in historical simulations of the same model in 1970-2000 (upper right corner in Figure 4). This means that this model yields a change in winter atmospheric circulation.

3 Discussion

The classification results of this study are based on a fairly simple use of a neural network, without any particular tuning (e.g. low number of layers of neurons, simple algorithm, no convolutional layer, etc.). For example, only one hidden layer of 256 neurons is sufficient to correctly classify summer SLP. The training sets (2000 days for each model) are small, compared to what is potentially available in the CMIP6 archive. This choice is made to avoid overfitting. The classification scores could be improved by considering larger training sets or tuning neural network parameters, although the dependence on the season would remain unchanged.

Using other climate fields (temperature, precipitation, wind speed) would be possible, but since they highly depend on the model horizontal resolution and other idiosyncrasies like local biases, it is expected that model classifiability could be enhanced in a fairly trivial way. Hence, this type of analysis would be interesting to test the effects of multi-variate bias correction on a CMIP6 ensemble and reanalysis data. [18] used this strategy to design bias correction models, where a bias-correction model is trained so that a classifier is not able to recognize a bias corrected field from a reanalysis field.

It has been argued that internal climate variability is a major source of uncertainty for climate assessments [19, 20], as members of ensemble simulations could show different behavior. Such studies hence suggest that simulations of the same climate model should be treated separately, as their variability might differ. Our study brings a nuance by showing that simulation ensembles of the same model do yield an identifiable consistency, while the role of internal variability is more important in the winter season, which leads to a lack of classification skill.

4 Conclusion

We outlined the intriguing faculty of identifying a climate model from one daily SLP map, with a simple neural network. This property is essentially valid for the summer season. The 16 reference training models we considered are not only different from each other, but they are also different from the ERA5 reanalysis on the present-day period. It is much more difficult to differentiate the SLP from models and reanalysis for other seasons, which is also an interesting feature, because one can pool models in order to increase the sampling of data in a meaningful way.

This rather simple approach also allows us to identify the sorority of climate models. This means that different flavors of the same model (increasing resolution, changing land-use schemes, or even changing the ocean model) lead to similar atmospheric properties in the midlatitude regions (at least the North Atlantic), and hence shows the robustness of the atmospheric part of coupled models, especially in the summer.

AI-based weather forecast systems have been trained on the ERA5 reanalysis [8, 21]. It is tempting to use a larger data base for training in order to have a better sampling of interannual atmospheric variability [7]. This study shows that it would be misleading to take any climate model, because of the different spatial probability distributions of atmospheric variables between models, and it would require complex bias correction. This study has outlined good candidate GCMs (with large ensembles, similarities with ERA5) for this exercise.

Conversely, this also implies that the statistical properties of the atmospheric circulation from large ensemble model simulations (with the same model [22]) can be enriched with variations around the same model, provided that the horizontal resolutions are similar. Therefore, it makes sense to "pool" climate model simulations of the same family, to perform extreme event attribution analyses.

5 Data and Methods

5.1 Data

This study focuses on North Atlantic [50W–20E; 30N–65N] daily SLP fields from the CMIP6 [9] archive. The SLP fields were extracted for all runs of global climate models (GCMs) that propose daily fields for historical and SSP simulations [15]. We restricted the analysis to the data that is available on the IPSL database server, which is a subset of the ESGF data https://esgf-node.ipsl.upmc.fr/projects/esgf-ipsl/. For simplicity, some models (e.g. with only one member) were not used in this study (black lines in Figure A1).

We focus on the North Atlantic region (SI Figure 1), although other regions of the world could have been used, because this is where climate models seem to yield a large consensus [23]. SLP patterns marginally depend on climate change (unlike many other climate fields) [24], although the patterns yield a seasonal cycle [16]. Hence investigating North Atlantic SLP classification is deemed a difficult endeavor.

For training, we chose 16 GCMs (out of 47) because they yield the largest number of runs (and more than 3), and reflect the diversity of atmospheric models, as climate

modeling groups can provide simulations with variations on the resolution, ocean model, etc. [1].

In the training set, we considered the first 2000 days (for each season) in historical simulations, from 1970 to 2000, which corresponds to ≈ 22 years (hence 1970 to 1992). We used the first simulations in lexicographic order for the training set. For validation, we used the remaining ≈ 8 years of the same simulation. The test set included historical simulations (1970–2000) from the third simulations in lexicographic order (this choice was arbitrary). Therefore, the test sets (in Figures 2 to 4) are disjoint from the learning set.

For all models, the SLP fields were normalized by their average. Therefore, models cannot be identified by a potential bias in the mean, which is difficult to identify. The spatial standard deviations are rather similar across models (SI Figure 2).

All SLP fields (CMIP6 and ERA5) are interpolated onto $1^\circ \times 1^\circ$ degree maps. Therefore information on the horizontal resolution of each model is not used in the classification.

The tests sets include other runs from the same model, runs from other models (black lines in Figure A1), and scenario runs until the end of the 21st century. In this paper, we focused on the SSP5-8.5 (although computations were done for all SSPs [25]). This scenario might not be the most relevant for society [26], but it allows a higher signal-to-noise ratio, to identify responses of climate change.

5.2 Methods

The AI classification model we use is a basic method of image classification [10, 27], that can be used to identify clothes from pictures [12].

The AI model is a simple dense neural network with a single hidden layer of 256 neurons with relu activation functions and an output layer of 17 neurons (the maximum number of GCMs) with a softmax activation functions. The dense neural network use as inputs SLP fields flattened into vectors of $70 \times 36 = 2520$ values.

The neural network model (AI model) is trained over the training set constructed from models listed in Table A1. 5 epochs are used in the training and sparse categorical cross-entropy is optimized, with an adam optimizer. The training is repeated 20 times, because this procedure yields random selections of input data and weight initialization. At each new training step, we compute the score rate, i.e., the probability of correctly classifying an SLP field on a validation set, which is different from the training set. The scores of this procedure are shown in Figure 1, for validation sets that consist of other historical simulations. For each season, we keep the AI model that yields the highest score.

Those AI models are used to classify models that are *not* listed in Table A1. This helps identifying potential "sororities", when GCMs share common atmospheric components.

Then the AI models (for historical periods) are used to classify scenario simulations (SSP5-8.5) to track potential drifts in daily SLP patterns.

In Figure 1 we consider an arbitrary score probability of $p = 0.6 > 0.5$, which is higher than a "coin toss" probability and much higher than an equiprobable classification value over the 17 models of $p = 1/17 \approx 0.06$.

6 Figures

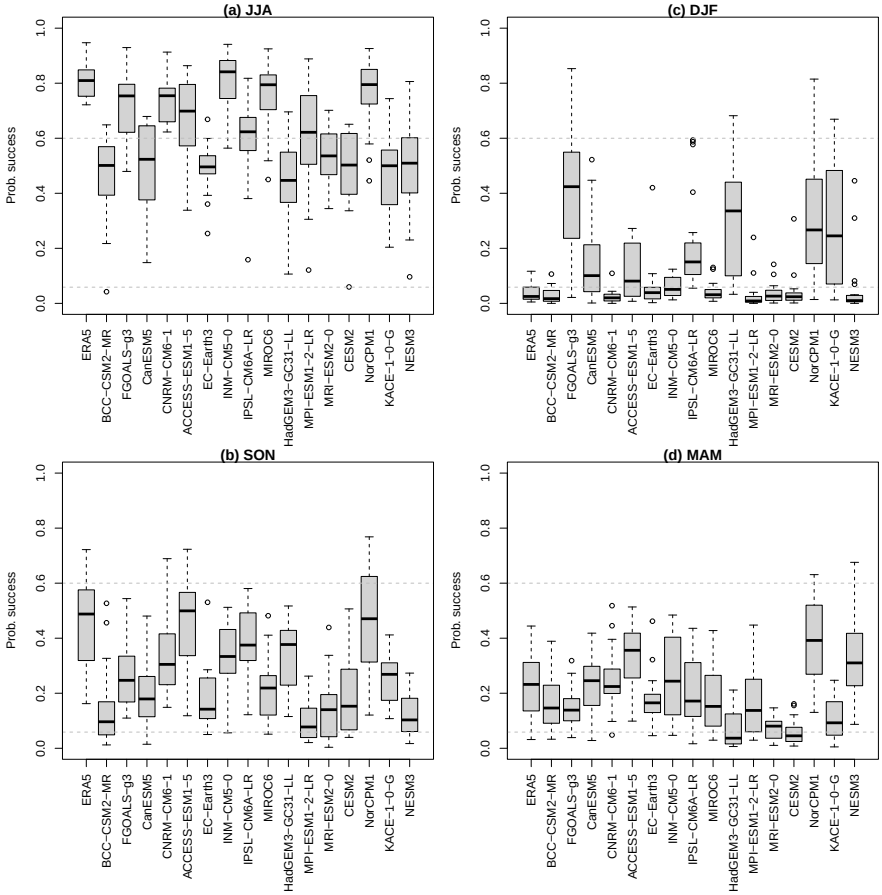

Fig. 1 Empirical probabilities that 16 CMIP6 models or ERA5 are correctly classified for each season (panels a to d). 20 training classifications were performed. The boxplots reflect the variability of the score probabilities of the training procedure (20 classifications). The horizontal dashed line represents the threshold probabilities $p = 0.6$ and $p = 1/17$ (a uniformly random classification).

Supplementary information. This article comes with supplementary information, with a table of all CMIP6 models that are considered, and summary figures.

Acknowledgements. Some ideas of this study emerged from discussions with colleagues at Météo-France (Enora Cariou, Julien Cattiaux and Aurélien Ribes). We thank the IPSL ESPRI group for maintaining the CMIP6 archive. Most computations were done with the IPSL GPU server.

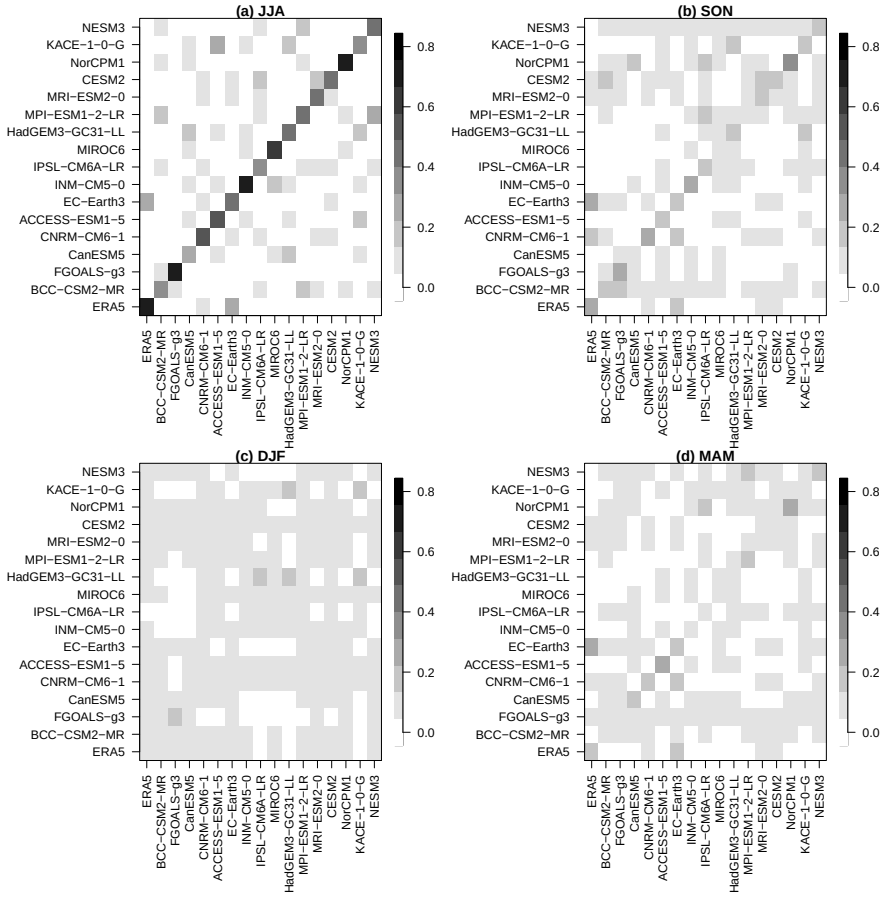

Fig. 2 Classification test. Empirical probabilities of classifying CMIP6 models in Table A1 or ERA5 onto those models, for the four seasons (panels a to d). Darker colors indicate higher probabilities. In those diagrams, the sum of probabilities along lines is 1. Model simulations listed on the vertical axis are classified onto the list of models on the horizontal axis. In this figure, the two lists of names coincide.

Declarations

- Funding: This work received the support of the grant ANR-20-CE01-0008-01 (SAMPRACE) and from Agence Nationale de la Recherche - France 2030, as part of the PEPR TRACCS programme under ANR-22-EXTR-0002. This work also received support from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101003469 (XAIDA).
- The author declare no competing interests.
- Data availability: ERA5 data was extracted from the climate explorer (https://climexp.knmi.nl/). CMIP6 data is available on ESGF nodes (e.g. https://esgf-node.ipsl.upmc.fr/projects/esgf-ipsl/). We used the data that is available on the IPSL data server.

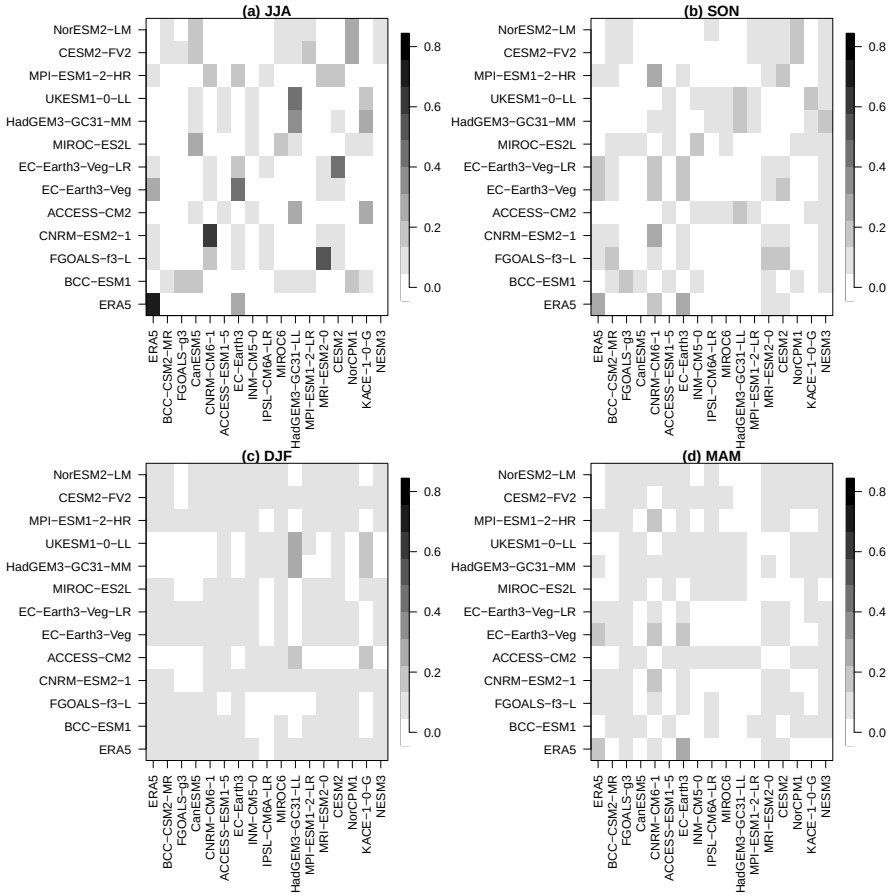

Fig. 3 Sorority classification. Empirical probabilities of classifying CMIP6 models (blue lines in Fig. A1 onto CMIP6 models in Table A1 or ERA5, for the four seasons (panels a to d). Darker colors indicate higher probabilities. Model simulations listed on the vertical axis are classified onto the list of models on the horizontal axis.

- Code availability: the codes for preparing CMIP6 files, and classification are available on from https://github.com/pascalyiou/CMIP6_CNN-Classification.git.
- Author contribution: PY conceived and performed the analyses, and wrote the manuscript. ST advised on the neural network implementation and contributed to the writing of the manuscript.

References

[1] Intergovernmental Panel On Climate Change (Ipcc). in *Annex II: Models* 1 edn, (ed.Gutiérrez, J M., A.-M. Tréguier) *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 2087–2138 (Cambridge University Press, 2023). URL https://www.cambridge.org/core/product/identifier/

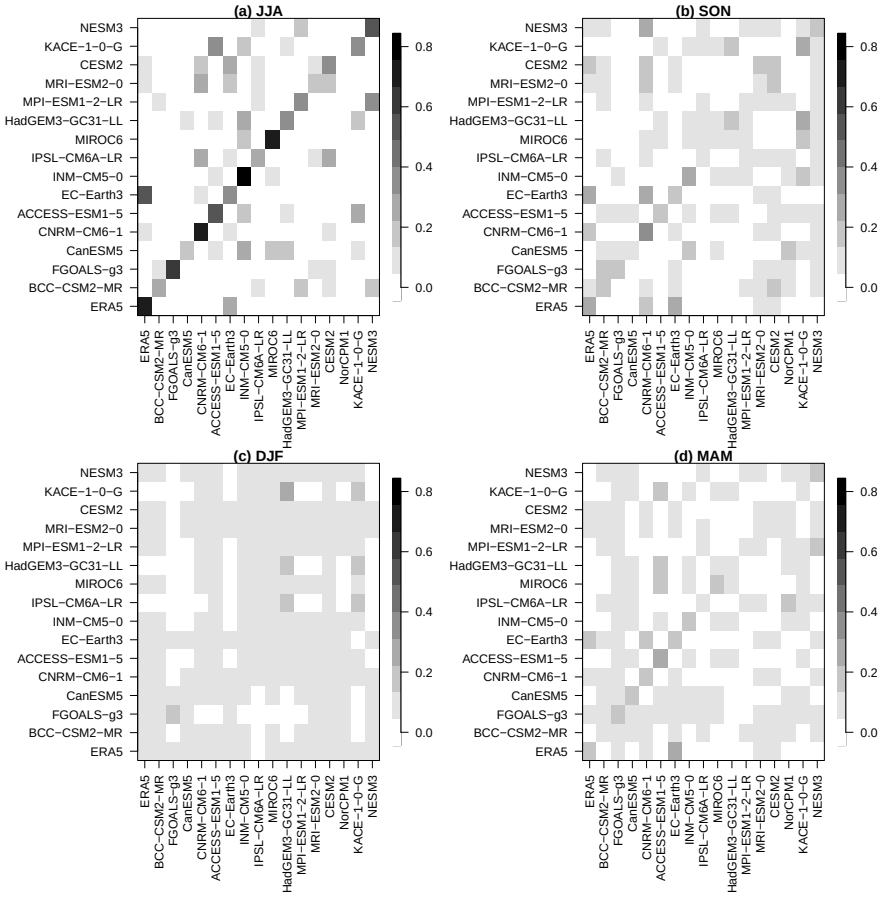

Fig. 4 Climate change classification. Empirical probabilities of classifying CMIP6 models in Table A1 or ERA5 for SSP5-8.5 simulations (2070–2100) onto those models (historical period: 1970–2000), for the four seasons (panels a to d). Darker colors indicate higher probabilities. Model SSP5-8.5 simulations listed on the vertical axis are classified onto historical simulations listed on the on the horizontal axis.

9781009157896/type/book.

[2] Kochkov, D. *et al.* Neural general circulation models for weather and climate. *Nature* **632**, 1060–1066 (2024). ISBN: 0028-0836 Publisher: Nature Publishing Group UK London.

[3] National Academies of Sciences Engineering and Medicine (ed.) *Attribution of Extreme Weather Events in the Context of Climate Change* (The National Academies Press, Washington, DC, 2016). URL www.nap.edu/catalog/21852/attribution-of-extreme-weather-events-in-the-context-of-climate-change.

[4] Vrac, M. Multivariate bias adjustment of high-dimensional climate simulations: the Rank Resampling for Distributions and Dependences (R 2 D 2) bias

correction. *Hydrology and Earth System Sciences* **22**, 3175 (2018).

[5] Robin, Y., Vrac, M., Naveau, P. & Yiou, P. Multivariate stochastic bias corrections with optimal transport. *Hydrology and Earth System Sciences Discussions* **2018**, 1–25 (2018). URL https://www.hydrol-earth-syst-sci-discuss.net/hess-2018-281/.

[6] Anwar, H. *et al.* Intercomparison of deep learning models in predicting streamflow patterns: insight from CMIP6. *Scientific Reports* **14**, 17468 (2024). URL https://www.nature.com/articles/s41598-024-63989-7.

[7] Huang, B., Liu, Z., Duan, Q., Rajib, A. & Yin, J. Unsupervised deep learning bias correction of CMIP6 global ensemble precipitation predictions with cycle generative adversarial network. *Environmental Research Letters* **19**, 094003 (2024). URL https://iopscience.iop.org/article/10.1088/1748-9326/ad66e6.

[8] Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023). URL https://doi.org/10.1126/science.adi2336. Publisher: American Association for the Advancement of Science.

[9] Eyring, V. *et al.* Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development* **9**, 1937–1958 (2016).

[10] Murphy, K. P. *Probabilistic machine learning: an introduction* Adaptive computation and machine learning (The MIT Press, Cambridge, Massachusetts London, England, 2022).

[11] Hersbach, H. *et al.* The ERA5 global reanalysis. *Quat. J. Roy. Met. Soc.* **146**, 1999–2049 (2020). ISBN: 0035-9009 Publisher: Wiley Online Library.

[12] Leithardt, V. Classifying garments from fashion-MNIST dataset through CNNs. *Advances in Science, Technology and Engineering Systems Journal* **6**, 989–994 (2021).

[13] Masson, D. & Knutti, R. Climate model genealogy: climate model genealogy. *Geophysical Research Letters* **38**, L08703 (2011). URL http://doi.wiley.com/10.1029/2011GL046864.

[14] Knutti, R., Masson, D. & Gettelman, A. Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters* **40**, 1194–1199 (2013). URL https://agupubs.onlinelibrary.wiley.com/doi/10.1002/grl.50256.

[15] Riahi, K. *et al.* The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global environmental change* **42**, 153–168 (2017). ISBN: 0959-3780 Publisher: Elsevier.

[16] Vrac, M., Vaittinada Ayar, P. & Yiou, P. Trends and variability of seasonal weather regimes. *International Journal of Climatology* **34**, 472–480 (2014). URL https://rmets.onlinelibrary.wiley.com/doi/10.1002/joc.3700.

[17] Döscher, R. *et al.* The EC-Earth3 Earth system model for the Coupled Model Intercomparison Project 6. *Geoscientific Model Development* **15**, 2973–3020 (2022). URL https://gmd.copernicus.org/articles/15/2973/2022/.

[18] François, B., Thao, S. & Vrac, M. Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks. *Climate Dynamics* **57**, 3323–3353 (2021). URL https://link.springer.com/10.1007/s00382-021-05869-8.

[19] Hawkins, E. & Sutton, R. The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society* **90**, 1095–1108 (2009). URL https://journals.ametsoc.org/doi/10.1175/2009BAMS2607.1.

[20] Deser, C., Terray, L. & Phillips, A. S. Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications. *Journal of Climate* **29**, 2237–2258 (2016). ISBN: 0894-8755 Publisher: American Meteorological Society.

[21] Price, I. *et al.* Probabilistic weather forecasting with machine learning. *Nature* (2024). URL https://www.nature.com/articles/s41586-024-08252-9.

[22] Bevacqua, E. *et al.* Advancing research on compound weather and climate events via large ensemble model simulations. *Nature Communications* **14**, 2145 (2023). ISBN: 2041-1723 Publisher: Nature Publishing Group UK London.

[23] Fernandez-Granja, J. A., Casanueva, A., Bedia, J. & Fernandez, J. Improved atmospheric circulation over Europe by the new generation of CMIP6 earth system models. *Climate Dynamics* **56**, 3527–3540 (2021). URL https://link.springer.com/10.1007/s00382-021-05652-9.

[24] Cusinato, E., Rubino, A. & Zanchettin, D. Winter Euro-Atlantic Climate Modes: Future Scenarios From a CMIP6 Multi-Model Ensemble. *Geophysical Research Letters* **48**, e2021GL094532 (2021). URL https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021GL094532.

[25] Meinshausen, M. *et al.* The shared socio-economic pathway (SSP) greenhouse gas concentrations and their extensions to 2500. *Geoscientific Model Development* **13**, 3571–3605 (2020). URL https://gmd.copernicus.org/articles/13/3571/2020/.

[26] Meinshausen, M. *et al.* A perspective on the next generation of Earth system model scenarios: towards representative emission pathways (REPs). *Geoscientific Model Development Discussions* **2023**, 1–40 (2023). ISBN: 1991-962X Publisher: Göttingen, Germany.

[27] Szeliski, R. in *Deep Learning* 187–271 (Springer International Publishing, Cham, 2022). URL https://link.springer.com/10.1007/978-3-030-34372-9_5. Series Title: Texts in Computer Science.

Appendix A List of CMIP6 models for training and testing

Table A1 Reference CMIP6 models that are considered for the classification of SLP. The ordering follows the research group name (2nd column). The training runs are the first of available runs (in lexicographic order). The test runs are the third in lexicographic order.

Model name	Group name	Training run	Test run
BCC-CSM2-MR	BCC	r1i1p1f1	r3i1p1f1
FGOALS-g3	CAS	r1i1p1f1	r4i1p1f1
CanESM5	CCCma	r1i1p1f1	r2i1p1f1
CNRM-CM6-1	CNRM-CERFACS	r1i1p1f2	r3i1p1f2
ACCESS-ESM1-5	CSIRO	r1i1p1f1	r3i1p1f1
EC-Earth3	EC-Earth-Consortium	r1i1p1f1	r4i1p1f1
INM-CM5-0	INM	r1i1p1f1	r3i1p1f1
IPSL-CM6A-LR	IPSL	r1i1p1f1	r3i1p1f1
MIROC6	MIROC	r1i1p1f1	r3i1p1f1
HadGEM3-GC31-LL	MOHC	r1i1p1f3	r3i1p1f3
MPI-ESM1-2-LR	MPI-M	r1i1p1f1	r3i1p1f1
MRI-ESM2-0	MRI	r1i1p1f1	r2i1p1f1
CESM2	NCAR	r1i1p1f1	r3i1p1f1
NorCPM1	NCC	r1i1p1f1	r4i1p1f1
KACE-1-0-G	NIMS-KMA	r1i1p1f1	r3i1p1f1
NESM3	NUIST	r1i1p1f1	r3i1p1f1

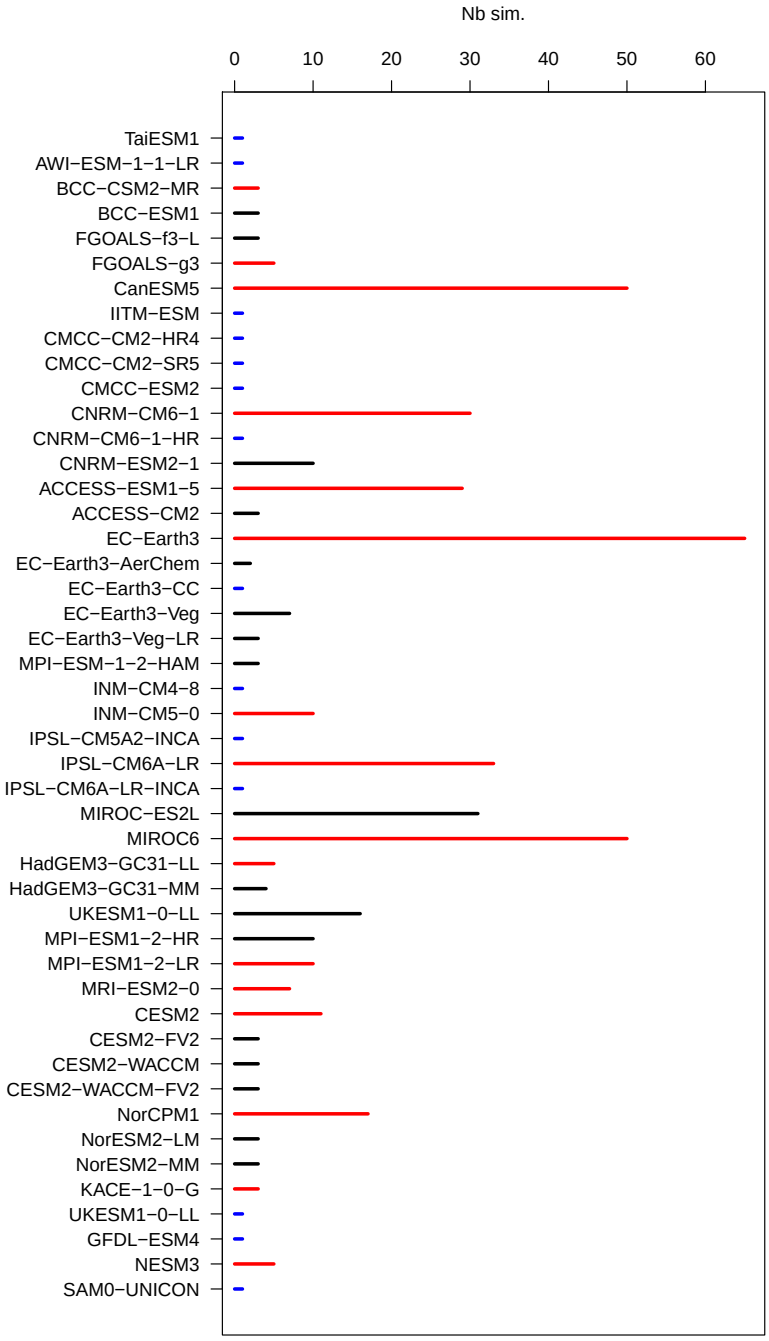

Fig. A1 List of CMIP6 models that are available on the IPSL computing server (horizontal axis) and number of runs per model. The red lines indicate the reference models that are used in this study (in Table A1). The black lines are for models with more than 2 simulations and that are used in the "sorority" experiments. The blue lines are for models with 1 simulation and are not used in this paper. The run of UKESM1-0-LL (bottom of the figure) was run by the NIMS-KMA group (who produced the KACE-1-0-G runs).