Tellus A
Dynamic Meteorology and Oceanography

Selecting CMIP6 Models for Future Arctic Storylines Using a Novel Performance Score

ORIGINAL RESEARCH PAPER

LISE SELAND GRAFF
OSKAR A. LANDGREN
KAJSA M. PARDING
XAVIER LEVINE
RYAN S. WILLIAMS
PRISCILLA A. MOONEY

*Author affiliations can be found in the back matter of this article

STOCKHOLM UNIVERSITY PRESS

ABSTRACT

Storylines are physically plausible scenarios of future climate change, statistically derived from an ensemble of climate model projections and organized according to the magnitude of projected changes in two or more remote drivers that strongly influence the spatial pattern of the climate response. Here, we provide novel insights into the Arctic storylines identified by Levine *et al.* (2024), where Barents-Kara Sea warming and lower-tropospheric Arctic warming during the extended summer season (May–October) were remote drivers, as we identify a set of models from the Coupled Model Intercomparison Project phase 6 to represent the storylines. We do this by first identifying models that are similar to these storylines in terms of each remote driver response and quantifying this similarity. Second, we evaluate the model's performance in terms of a simple performance score based on the mean normalized root-mean-square error for multiple climate variables of importance for the storylines. The normalized values vary between 0 and 1 for all variables, allowing them to exert a comparable influence on the score. The advantage of the score is that it provides an easily implementable and interpretable way of identifying models that are characterized by large errors relative to the rest of the ensemble. Finally, we combine the similarity estimate and the score to select models to represent the storylines. We focus on the Arctic during the extended summer season for which the storylines were designed, but also consider other seasons and regions. Through this exercise, we also document the methodology, benefits, and limitations of the score.

CORRESPONDING AUTHOR:
Lise Seland Graff
Norwegian Meteorological Institute, Oslo, Norway
lise.s.graff@met.no

KEYWORDS:
RMSE-based relative performance score; model selection; model evaluation; Arctic storylines; CMIP6

TO CITE THIS ARTICLE:
Graff, L.S., Landgren, O.A., Parding, K.M., Levine, X., Williams, R.S. and Mooney, P.A. (2026) Selecting CMIP6 Models for Future Arctic Storylines Using a Novel Performance Score. *Tellus A: Dynamic Meteorology and Oceanography* 78(1): 1–21. DOI: https://doi.org/10.16993/tellusa.4111

1 INTRODUCTION

Any application of modeled climate data should be supported by an evaluation of the basic performance of the model, that is, how well it simulates the observed climate, for the variables, regions, and seasons of interest. When using multi-model ensembles such as the global climate model simulations from the Coupled Model Intercomparison Project phase 6 (CMIP6; Eyring *et al.*, 2016a), it can be necessary to identify models that do not perform as well as the rest of the ensemble and that warrant additional investigation or even exclusion from further analysis. This is particularly relevant when it is not desirable or even permissible to use the entire ensemble of models and a subset must be chosen instead. An example is dynamical downscaling of CMIP6 models, as it is computationally prohibitive to consider downscaling the full CMIP6 ensemble with a regional climate model. Another is identifying individual CMIP6 models that are most representative of specific storylines of future climate change (Williams *et al.*, 2024). While the final model selection tends to be subjective, the process can benefit from being guided by objective evaluations.

A number of methods and tools for evaluating the simulated climate of multi-model ensembles already exist (e.g., Eyring *et al.*, 2019), ranging in complexity from simple online applications like GCMeval (Parding *et al.*, 2020) to more comprehensive evaluation systems like the Earth System Model Evaluation tool (ESMValTool; Eyring *et al.*, 2016b; Righi *et al.*, 2020). Methods can be based on basic statistical metrics like mean relative error (e.g., Gleckler *et al.*, 2008), normalized error variance (e.g., Reichler and Kim, 2008), and root-mean-square error (RMSE) that has been normalized by observational uncertainty (e.g., Karpechko *et al.*, 2010), or more complex frameworks like empirical orthogonal functions (e.g., Ashfaq *et al.*, 2022; Hannachi, 2021), common empirical orthogonal functions (e.g., Benestad *et al.*, 2023; Hannachi *et al.*, 2022; Hannachi, 2021; Sengupta and Boyle, 1998), atmospheric circulation-type statistics (e.g., Brands, 2022), or clustering (e.g., Samantaray *et al.*, 2023; Yokoi *et al.*, 2011). Furthermore, there are specialized diagnostics like the Taylor diagram (Taylor, 2001), which can summarize multiple metrics in one single plot. Each of these scores and diagnostics offers different valuable insights into the climate models and the ensembles, but each also has its limitations. Condensing model evaluation into one single metric remains challenging, although great progress has been made to combine different diagnostics to produce a single-value performance score for each model (e.g., Gleckler *et al.*, 2008; Hu *et al.*, 2022; Karpechko *et al.*, 2010; Reichler and Kim, 2008; Samantaray *et al.*, 2023).

Here, we define a novel and simple score for comparing data from multiple climate models to reference data from reanalysis and gridded observations that is based on average normalized RMSE (NRMSE) for multiple variables, *the RMSE-based relative performance score (RRPS)*. Unlike previous studies, we normalize the RMSE values so that they vary between 0 and 1 for all variables. This simplifies the interpretation of the results as it allows for all variables to exert a comparable influence on the score. By design, the score provides a measure of the *relative* model performance within the ensemble and facilitates identifying models that stand out from the rest of the ensemble.

We demonstrate the benefits and limitations of the score by using it to aid the selection of a single CMIP6 model or a set of CMIP6 models to represent storylines of Arctic climate change previously identified by Levine *et al.* (2024). Storylines can be defined as physically plausible potential pathways of future climate change (Zappa and Shepherd, 2017). The storylines methodology provides a way of examining several possible future outcomes with distinctly different climatological characteristics in parallel. Levine *et al.* (2024) showed that during the extended summer season (May, June, July, August, September, and October; MJJASO), large parts of the Arctic inter-model spread can be explained by inter-model differences in two predictors: Barents-Kara Sea surface warming and Arctic amplification. From this, they derived four storylines of summer Arctic climate based on the strength (weak or strong) of these predictors.

To establish a set of CMIP6 models to represent each of the four storylines defined in Levine *et al.* (2024), we first identify a set of models that are similar to the storylines in terms of their Barents-Kara Sea warming and Arctic amplification and compute an estimate of this similarity. Then we use the score to evaluate the relative present-day performance of the models considering four impact-relevant variables used in Levine *et al.* (2024): near-surface temperature (*tas*), total (large-scale and convective) precipitation rate (*pr*), 850-hPa air temperature (*ta*850), and 850-hPa zonal wind (*ua*850). Finally, we identify the models most suitable for representing the storylines by combining the similarity estimate and the scores.

In what follows, we first provide an overview of the data used in Section 2 and a detailed description of the RRPS methodology in Section 3. We summarize pertinent information from Levine *et al.* (2024), introduce a similarity estimate, and use this estimate to establish a list of candidate models for each storyline in Section 4.1. We then consider the similarity estimate in combination with scores for Arctic extended summer in section 4.2 and in combination with scores for a more comprehensive set of seasons and regions in Section 4.3. Finally, we summarize and discuss our findings in Section 5.

2 DATA

We use data from CMIP6 historical experiments (Eyring *et al.*, 2016a) from 50 models (see appendix A for a complete list). The models were selected based on: (1) whether they had performed the CMIP6 historical and the shared socioeconomic pathway scenario that corresponds to a 8.5 W m^{-1} increase in radiative forcing by the end of the 21st century (SSP5-8.5; O'Neill *et al.*, 2016); and (2) whether all necessary fields were available at the Earth System Grid Federation (ESGF) at the time of the data retrieval: monthly *tas*, *pr*, zonal wind (*ua*), air temperature (*ta*), and surface pressure (*ps*). For *ua* and *ta*, we extract the 850-hPa level (*ua*850 and *ta*850). While *ps* is not explicitly included in the score, it is used to filter out below-surface grid points from the 850-hPa fields for consistency across the models (this is needed as some models extrapolate and others assign missing/fill values to sub-surface grid points). Hence, we consider four variables for the score: *tas*, *pr*, *ua*850, and *ta*850.

To verify the models, we use reference data from the 5th generation of the European Center for Medium-Range Weather Forecasts' Reanalysis (ERA5; Hersbach *et al.*, 2020) for *ps*, *tas*, *ta*850, and *ua*850 (Hersbach *et al.*, 2019a,b) and from the Global Precipitation Climatology Project precipitation analysis (GPCP; Adler *et al.*, 2003) for *pr* (Adler *et al.*, 2016). We use years 1985–2014 in all cases.

3 THE RRPS METHODOLOGY

The RRPS quantifies the model's ability to simulate the historical climate for the selected variables as a single value, and is defined as the average NRMSE taken over a set of variables. We base the score on RMSE as it is not affected by cancellation of positive and negative errors (like the bias) and emphasizes large errors. Here, the score is based on *tas*, *pr*, *ta*850, and *ua*850. We consider both annual and seasonal scores. Below, we outline the procedure for computing the scores step by step.

1. As an initial step, we bilinearly interpolate all data to a common 1° × 1° grid in the horizontal.
2. We then compute monthly climatologies, that is, multi-year mean values computed separately for each month of the year.
3. To ensure consistency across the models, we mask out sub-surface grid points (i.e., in areas of high elevation) in *ua*850 and *ta*850 using *ps*.
4. The RMSE values are computed as the root of the squared differences between the monthly climatological values from the CMIP6 models and the relevant reference data sets for the same time period, 1985–2014. For a variable *X*, model *c*, month *m*, and reference data set denoted *REF*, we compute the RMSE values separately for each zonal grid point *i* and meridional grid point *j*:

$$\mathrm{RMSE}_{X,c,m,i,j} = \sqrt{\left(X_{c,m,i,j} - X_{REF,m,i,j}\right)^2} \tag{1}$$

5. To account for variations of grid-cell area with latitude, RMSE values are spatially averaged using a standard cosine weighting:

$$\mathrm{RMSE}_{X,c,m} = \frac{\sum_{i,j} \cos(\phi_j) \times \mathrm{RMSE}_{X,c,m,i,j}}{\sum_{i,j} \cos(\phi_j)} \tag{2}$$

   where ϕ is latitude.
6. We average the RMSE values temporally, using all months for annual values or a subset for seasonal values, and then normalize them so that they vary between 0 (for the best model) and 1 (for the worst model) following the approach used by Ashfaq *et al.* (2022) to normalize the absolute error (their equation 1):

$$\mathrm{NRMSE}_{X,c} = \frac{\mathrm{RMSE}_{X,c} - min(\mathrm{RMSE}_{X,c_{all}})}{\max(\mathrm{RMSE}_{X,c_{all}}) - \min(\mathrm{RMSE}_{X,c_{all}})} \tag{3}$$

   where the subscript c_{all} indicates that all models are used.
7. Finally, the RRPS for each model (RRPS$_c$) is computed as the average NRMSE across the variables *X*:

$$\mathrm{RRPS}_c = \frac{1}{N} \sum_{X=1}^{X=N} \mathrm{NRMSE}_{X,c} \tag{4}$$

An advantage of this normalization technique is that the NRMSE values vary within the same range (0 to 1) for all variables, meaning that the importance of all variables is more comparable than if we use the non-normalized RMSE. The score has the same range as the NRMSE, with low scores being favorable, indicating smaller errors than the other models, and higher values indicative of larger errors. The top score of 0 is only obtained if the same model is consistently the best performing model (NRMSE = 0) for all variables and the worst score of 1 is only obtained if the same model has the largest relative errors in the ensemble (NRMSE = 1) for all variables. Another advantage of the normalization is that it easily allows for different variables to be assigned weights if necessary.

We compute the score separately for six geographical regions: the whole globe, the tropics (15°S–15°N), Northern Hemisphere (NH) mid-latitudes (15°N–55°N), Southern Hemisphere (SH) mid-latitudes (15°S–55°S), Arctic (55°N–90°N), and Antarctic (55°S–90°S). One should note that our somewhat broad definition of the Arctic is chosen for consistency with Levine *et al.* (2024) and includes large land areas with boreal forest, which have a significantly different climate (e.g., more precipitation, warmer summers) than the ocean- and sea-ice-dominated central Arctic. For both the Arctic and Antarctic, the placement of the equatorward border means that it better captures the jet streams and storm tracks than other more restrictive definitions of the polar regions.

We consider the annual and seasonal scores. Our main focus is on the extended NH warm season (MJJASO), as this season was used to construct the storylines in Levine *et al*. (2024). However, we also include the four standard three-month seasons: December, January, and February (DJF), March, April, and May (MAM), June, July, and August (JJA), and September, October, and November (SON).

Next, we identify candidate models from the CMIP6 ensemble to represent the storylines from Levine *et al*. (2024) and provide an estimate of how similar the candidate models are to the storylines. Then, we combine these similarity estimates with the scores to produce an estimate of the overall fit for each candidate model, and use this as a basis for selecting models to represent the storylines, first for Arctic MJJASO (Section 4.2) and then for the full set of regions and seasons (Section 4.3).

4 RESULTS

4.1 ARCTIC STORYLINES OF CLIMATE CHANGE

Storylines are physically plausible scenarios of future climate change, statistically derived from an ensemble of climate model projections, which are organized according to the magnitude of the projected changes in two or more remote drivers (e.g., Arctic amplification) that strongly influence the spatial pattern of the climate response (Zappa and Shepherd, 2017). Levine *et al*. (2024) applied the storylines methodology to find potential future pathways of Arctic climate change based on the CMIP6 historical and SSP5-8.5 scenario from an ensemble of CMIP6 models. An extended NH warm season MJJASO was defined because of its importance for societal and ecological impacts of climate change, that is, Arctic wildfires and permafrost thaw, which are especially pronounced during the summer (e.g., Chadburn *et al*., 2017; Masrur *et al*., 2018; McCarty *et al*., 2021). Using a multivariate linear regression framework, Levine *et al*. (2024) regressed the pattern of change in *tas*, *ua*850, *pr*, and sea-ice fraction onto two predictors that were found to explain most of the inter-model variability: Barents-Kara Sea warming and lower-tropospheric Arctic warming. Based on this, four storylines of future Arctic climate change were identified (see also Table 1): storyline A, *weak* Arctic amplification and *strong* Barents-Kara Sea warming; storyline B, *strong* Arctic Amplification and *strong* Barents-Kara Sea warming; storyline C, *weak* Arctic amplification and *weak* Barents-Kara Sea warming; and storyline D, *strong* Arctic Amplification and *weak* Barents-Kara Sea warming.

Building on the work of Levine *et al*. (2024), we identify models from the CMIP6 ensemble that are representative of the different storylines. We start by evaluating the model's ability to represent the Arctic storylines in terms of their projected strengthening of Arctic amplification and surface warming of the Barents-Kara seas (Figure 1). For each storyline (blue and red dots in Figure 1), we define a region around it, constrained by the isoline where the Euclidean distance from the storyline point is 0.75 in predictor space, and consider models that fall within this region to be candidate models for that storyline. This threshold is selected to ensure that the candidate models are close to the storyline point and that all storylines have multiple candidate models. For models with more than one realization, we consider the realization that is closest to the storyline point. A detailed overview of the candidate models and realizations we use for the different storylines is provided in Table 1.

Having defined a set of candidate models for each storyline, we next quantify the representativeness (or similarity) of the candidate models by computing the Euclidean distance between the location of the candidate models and the relevant storyline point (Figure 1). The Euclidean distances for all candidate models are given in Table 1. Results show that storylines B, C, and D all have at least one candidate model that is very close to the storyline point, with the closest models found at an Euclidean distance of 0.12 (MIROC-ES2L), 0.09 (CESM2-WACCM), and 0.16 (NorESM2-MM) from the storyline points. For storyline A, the four candidate models are all found somewhat further away, with the smallest distance being 0.38 (CNRM-CM6-1).

In what follows, we use the score described in Section 3 to assess the performance of the full set of CMIP6 models (Appendix A), with particular emphasis on the performance of candidate models relative to the other models in the ensemble. We then use the scores in combination with the Euclidean distance to select a single model or a set of models to represent each storyline. We start by considering the same region and season as Levine *et al*. (2024), namely the Arctic during MJJASO in Section 4.2.

4.2 ARCTIC EXTENDED SUMMER

The Arctic MJJASO (non-normalized) RMSE values for *tas*, *pr*, *ua*850, and *ta*850 are shown in Figure 2. While most values are within one standard deviation of the multi-model mean (white cells), others deviate more (blue/red cells). It is clear that some models are characterized by values that are favorable (low) relative to the rest of the multi-model ensemble (blue cells), while others are characterized by more disadvantageous (higher) values (red cells). In some cases, the RMSE values can deviate from the multi-model mean by several standard deviations. The multi-model statistics (yellow cells at the bottom of the table) show that

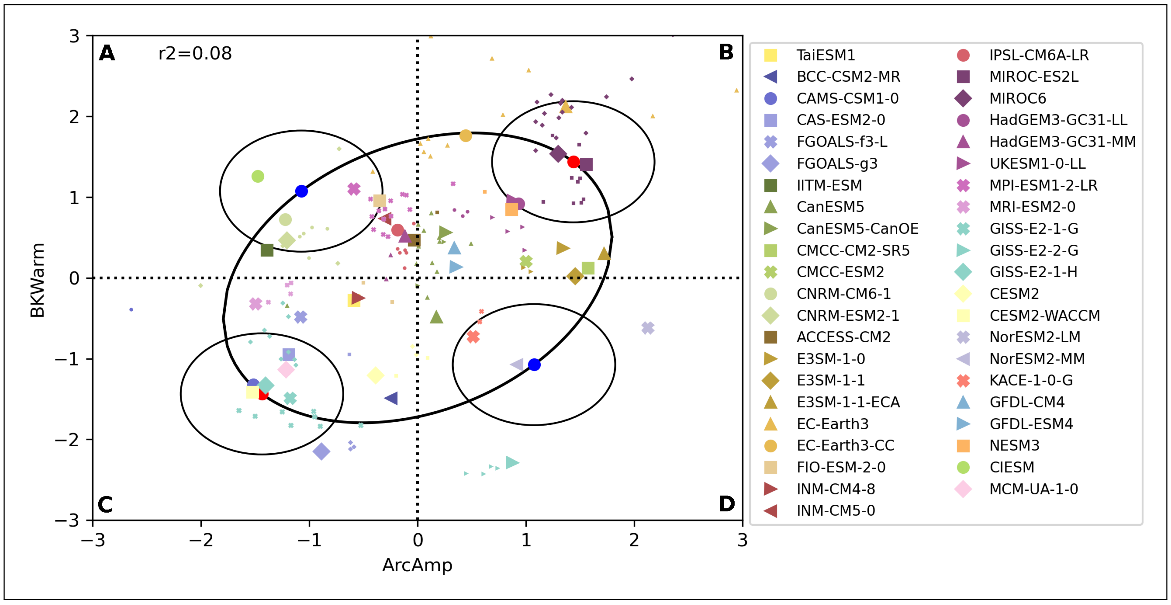
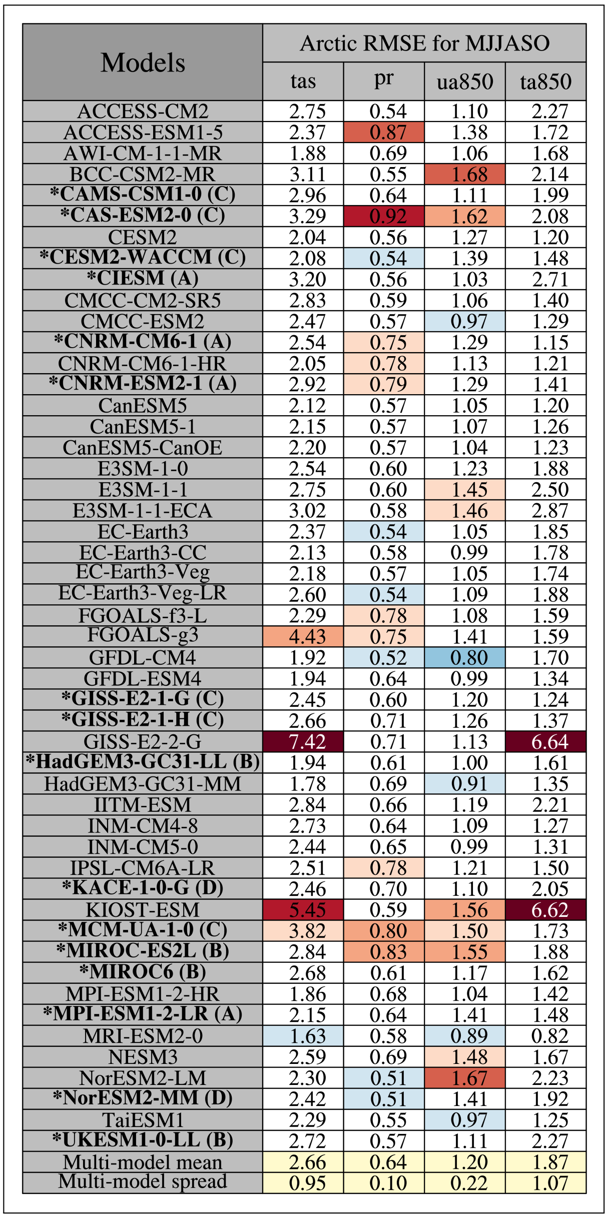


Figure 1 Predictor diagram for Arctic storylines showing the area-mean future change in Arctic *ta*850 (ArcAmp; x-axis) and Barents-Kara-Sea sea surface temperature (BKWarm; y-axis) for MJJASO for the models (legend) used in Levine *et al*. (2024). The future changes are computed as the difference between SSP5-8.5 (2070–2099) and the CMIP6 historical (1985–2014) normalized by the global-mean annual-mean change in *tas*. ArcAmp and BKWarm are relative to the multi-model mean and normalized by each model's standard deviation; hence, a predictor value of 1 means that it deviates from the multi-model mean by 1 standard deviation. The ellipse shows the 80% confidence region for the predictors and the blue and red dots on the ellipse shows the four storylines (A–D) as in Levine *et al*. (2024). The circles denote where the Euclidean distance from the storyline point is 0.75. For models with multiple realizations, markers with two different sizes are shown: the large marker shows the realization that is closest to the relevant storyline point; other realizations that were also considered by Levine *et al*. (2024) are shown as substantially smaller markers, but with the same color and marker shape.

STORYLINE	PREDICTOR 1	PREDICTOR 2	CANDIDATE MODEL	ABBR.	REALIZATION	ED
A	ArcAmp–	BKWarm+	CNRM-CM6-1	A1	r1i1p1f2	0.38
			CIESM	A2	r1i1p1f1	0.44
			MPI-ESM1-2-LR	A3	r17i1p1f1	0.48
			CNRM-ESM2-1	A4	r1i1p1f2	0.62
B	ArcAmp+	BKWarm+	MIROC-ES2L	B1	r8i1p1f2	0.12
			MIROC6	B2	r23i1p1f1	0.17
			HadGEM3-GC31-LL	B3	r3i1p1f3	0.73
			UKESM1-0-LL	B4	r8i1p1f2	0.74
C	ArcAmp–	BKWarm–	CESM2-WACCM	C1	r1i1p1f1	0.09
			GISS-E2-1-H	C2	r5i1p1f2	0.11
			CAMS-CSM1-0	C3	r1i1p1f1	0.14
			GISS-E2-1-G	C4	r3i1p5f1	0.26
			MCM-UA-1-0	C5	r1i1p1f2	0.38
			CAS-ESM2-0	C6	r3i1p1f1	0.55
D	ArcAmp+	BKWarm–	NorESM2-MM	D1	r1i1p1f1	0.16
			KACE-1-0-G	D2	r1i1p1f1	0.66

Table 1 Overview of Arctic storylines (A–D) and CMIP6 candidate models. The storylines are listed in column 1 and the two predictors the storylines are based on, the strength (+/–) of the Arctic amplification (ArcAmp+/–) relative to the multi-model mean, and the strength (+/–) of the Barents-Kara Sea warming (BKWarm+/–) relative to the multi-model mean, are listed in columns 2 and 3. The CMIP6 candidate models are listed in column 4, their abbreviations (abbr.) in column 5, the realization that is closest to the storyline point in the predictor-phase diagram (Figure 1) in column 6, and the Euclidean distance (ED) between the model/realization and the storyline point in the predictor-phase diagram in column 7. We omit FIO-ESM-2-1 (r1i1p1f1; ED = 0.73) from storyline A due to technical issues and EC-Earth3 (r112i1p1f1; ED = 0.69) from storyline B as *ps* was not available. Note that the models are ranked by the Euclidean distance for each storyline.



Models	Arctic RMSE for MJJASO			
	tas	pr	ua850	ta850
ACCESS-CM2	2.75	0.54	1.10	2.27
ACCESS-ESM1-5	2.37	0.87	1.38	1.72
AWI-CM-1-1-MR	1.88	0.69	1.06	1.68
BCC-CSM2-MR	3.11	0.55	1.68	2.14
***CAMS-CSM1-0 (C)**	2.96	0.64	1.11	1.99
***CAS-ESM2-0 (C)**	3.29	0.92	1.62	2.08
CESM2	2.04	0.56	1.27	1.20
***CESM2-WACCM (C)**	2.08	0.54	1.39	1.48
***CIESM (A)**	3.20	0.56	1.03	2.71
CMCC-CM2-SR5	2.83	0.59	1.06	1.40
CMCC-ESM2	2.47	0.57	0.97	1.29
***CNRM-CM6-1 (A)**	2.54	0.75	1.29	1.15
CNRM-CM6-1-HR	2.05	0.78	1.13	1.21
***CNRM-ESM2-1 (A)**	2.92	0.79	1.29	1.41
CanESM5	2.12	0.57	1.05	1.20
CanESM5-1	2.15	0.57	1.07	1.26
CanESM5-CanOE	2.20	0.57	1.04	1.23
E3SM-1-0	2.54	0.60	1.23	1.88
E3SM-1-1	2.75	0.60	1.45	2.50
E3SM-1-1-ECA	3.02	0.58	1.46	2.87
EC-Earth3	2.37	0.54	1.05	1.85
EC-Earth3-CC	2.13	0.58	0.99	1.78
EC-Earth3-Veg	2.18	0.57	1.05	1.74
EC-Earth3-Veg-LR	2.60	0.54	1.09	1.88
FGOALS-f3-L	2.29	0.78	1.08	1.59
FGOALS-g3	4.43	0.75	1.41	1.59
GFDL-CM4	1.92	0.52	0.80	1.70
GFDL-ESM4	1.94	0.64	0.99	1.34
***GISS-E2-1-G (C)**	2.45	0.60	1.20	1.24
***GISS-E2-1-H (C)**	2.66	0.71	1.26	1.37
GISS-E2-2-G	7.42	0.71	1.13	6.64
***HadGEM3-GC31-LL (B)**	1.94	0.61	1.00	1.61
HadGEM3-GC31-MM	1.78	0.69	0.91	1.35
IITM-ESM	2.84	0.66	1.19	2.21
INM-CM4-8	2.73	0.64	1.09	1.27
INM-CM5-0	2.44	0.65	0.99	1.31
IPSL-CM6A-LR	2.51	0.78	1.21	1.50
***KACE-1-0-G (D)**	2.46	0.70	1.10	2.05
KIOST-ESM	5.45	0.59	1.56	6.62
***MCM-UA-1-0 (C)**	3.82	0.80	1.50	1.73
***MIROC-ES2L (B)**	2.84	0.83	1.55	1.88
***MIROC6 (B)**	2.68	0.61	1.17	1.62
MPI-ESM1-2-HR	1.86	0.68	1.04	1.42
***MPI-ESM1-2-LR (A)**	2.15	0.64	1.41	1.48
MRI-ESM2-0	1.63	0.58	0.89	0.82
NESM3	2.59	0.69	1.48	1.67
NorESM2-LM	2.30	0.51	1.67	2.23
***NorESM2-MM (D)**	2.42	0.51	1.41	1.92
TaiESM1	2.29	0.55	0.97	1.25
***UKESM1-0-LL (B)**	2.72	0.57	1.11	2.27
Multi-model mean	2.66	0.64	1.20	1.87
Multi-model spread	0.95	0.10	0.22	1.07

Figure 2 Overview of models (sorted alphabetically; column 1) and Arctic MJJASO RMSE values for *tas* (K; column 2), *pr* (mm day^{-1}; column 3), *ua*850 (m s^{-1}; column 4), and *ta*850 (K; column 5). Cells with blue/red shading indicate that the RMSE values are lower/larger than the multi-model mean by one standard deviation or more, and the darker the shading the more the value deviates. The multi-model mean RMSE and spread (one standard deviation) is given in the two bottom rows (yellow shading). Candidate models for Arctic storylines are shown in bold with a preceding asterisk and with the storyline (A/B/C/D) given in a parenthesis following the model name.

the mean and spread vary from variable to variable, with some variables being characterized by a more pronounced spread relative to the multi-model mean than others. The diversity between variables emphasizes the importance of normalizing the RMSE values before combining them into a single score. This is not only needed to achieve dimensionless numbers, but also prevents the score from being dominated by the presence of variables that are characterized by larger errors than the others, thus ensuring that all four variables exert a more comparable influence on the score. The inclusion of two temperature variables means that the combined influence of temperature on the score will be larger than that from precipitation or zonal wind. However, we opt to include both temperature fields for consistency with Levine *et al*. (2024) and because of the importance of these variables for impact assessments, as identified in Levine *et al*. (2024).

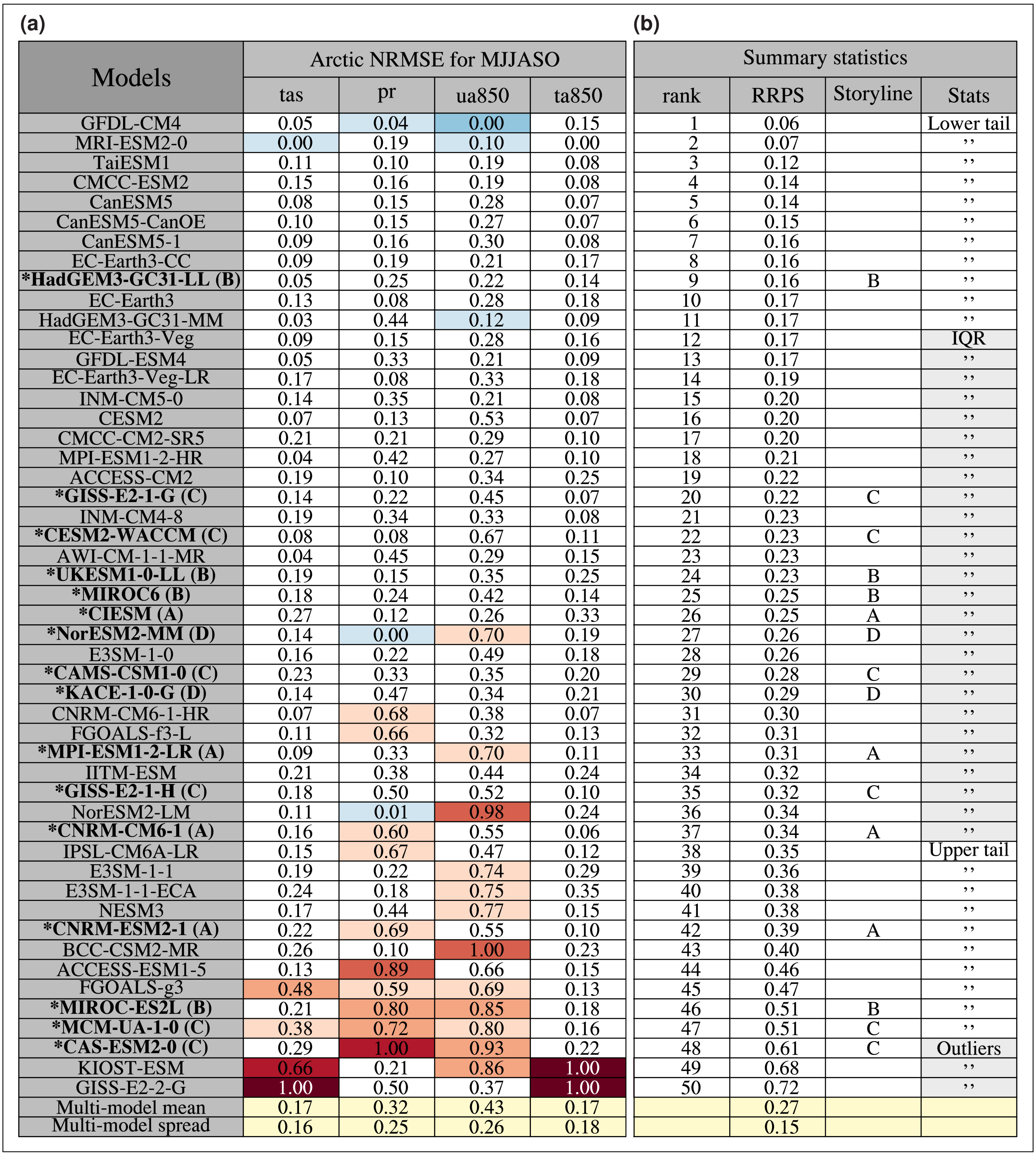
It is clear that models that perform substantially better than the multi-model mean in terms of Arctic MJJASO NRMSE (blue cells in Figure 3a) tend to perform well in terms of the score (b) and be placed in the upper part of the table, and vice versa. This is in line with the score being positively correlated with the NRMSE of the individual variables, with linear correlation coefficients of 0.79 for *tas*, 0.59 for *pr*, 0.73 for *ua*850, and 0.66 for *ta*850 (all values are significant at the $p = 0.05$ level following a Student's *t*-test). The Arctic MJJASO scores range between 0.06 and 0.72. The best score is close to 0, meaning that the best-performing model (GFDL-CM4) performs well for all variables, although it does not consistently have the lowest NRMSE (which would give a score of 0). The worst (highest) score, however, is quite a bit better than the theoretical maximum value of 1, meaning that the relative performance of the model in question varies more from variable to variable; that is, no single model has the poorest relative performance for all four variables.

Storylines A, B, and C all have candidate models within the upper tail of the distribution (Figure 3b), defined as the range between the upper quartile (Q2) and Q2 plus 1.5 times the inter-quartile range (IQR). Storyline C also has a candidate model that is defined as an outlier (score exceeding Q2 plus 1.5 times the IQR). Only a single candidate model, HadGEM3-GC31-LL (a candidate for storyline B), places within the lower tail, defined as the range between the first quartile (Q1) and Q1 minus 1.5 times the IQR. The remaining candidate models are within the IQR.

In line with the score being a measure of relative model performance and the goal being to identify models with large errors compared to the rest of the multi-model ensemble, we consider candidate models with scores that are within the upper tail (including outliers) to be less preferable based on the relative performance for the variables, region, and season of primary interest.

To determine which model is most suited to represent the four storylines, we consider the Arctic MJJASO scores in combination with the Euclidean distance between the models and relevant storyline point in predictor space (Figure 1; Table 1). Specifically, we use the product of these two quantities to produce a measure of the overall fit for each model:

$$fit = \mathrm{RRPS}_c(s, r) \times \mathrm{ED}_c(s = \mathrm{MJJASO}, r = \mathrm{Arctic}) \qquad (5)$$



(a)

Models	Arctic NRMSE for MJJASO			
	tas	pr	ua850	ta850
GFDL-CM4	0.05	0.04	0.00	0.15
MRI-ESM2-0	0.00	0.19	0.10	0.00
TaiESM1	0.11	0.10	0.19	0.08
CMCC-ESM2	0.15	0.16	0.19	0.08
CanESM5	0.08	0.15	0.28	0.07
CanESM5-CanOE	0.10	0.15	0.27	0.07
CanESM5-1	0.09	0.16	0.30	0.08
EC-Earth3-CC	0.09	0.19	0.21	0.17
***HadGEM3-GC31-LL (B)**	0.05	0.25	0.22	0.14
EC-Earth3	0.13	0.08	0.28	0.18
HadGEM3-GC31-MM	0.03	0.44	0.12	0.09
EC-Earth3-Veg	0.09	0.15	0.28	0.16
GFDL-ESM4	0.05	0.33	0.21	0.09
EC-Earth3-Veg-LR	0.17	0.08	0.33	0.18
INM-CM5-0	0.14	0.35	0.21	0.08
CESM2	0.07	0.13	0.53	0.07
CMCC-CM2-SR5	0.21	0.21	0.29	0.10
MPI-ESM1-2-HR	0.04	0.42	0.27	0.10
ACCESS-CM2	0.19	0.10	0.34	0.25
***GISS-E2-1-G (C)**	0.14	0.22	0.45	0.07
INM-CM4-8	0.19	0.34	0.33	0.08
***CESM2-WACCM (C)**	0.08	0.08	0.67	0.11
AWI-CM-1-1-MR	0.04	0.45	0.29	0.15
***UKESM1-0-LL (B)**	0.19	0.15	0.35	0.25
***MIROC6 (B)**	0.18	0.24	0.42	0.14
***CIESM (A)**	0.27	0.12	0.26	0.33
***NorESM2-MM (D)**	0.14	0.00	0.70	0.19
E3SM-1-0	0.16	0.22	0.49	0.18
***CAMS-CSM1-0 (C)**	0.23	0.33	0.35	0.20
***KACE-1-0-G (D)**	0.14	0.47	0.34	0.21
CNRM-CM6-1-HR	0.07	0.68	0.38	0.07
FGOALS-f3-L	0.11	0.66	0.32	0.13
***MPI-ESM1-2-LR (A)**	0.09	0.33	0.70	0.11
IITM-ESM	0.21	0.38	0.44	0.24
***GISS-E2-1-H (C)**	0.18	0.50	0.52	0.10
NorESM2-LM	0.11	0.01	0.98	0.24
***CNRM-CM6-1 (A)**	0.16	0.60	0.55	0.06
IPSL-CM6A-LR	0.15	0.67	0.47	0.12
E3SM-1-1	0.19	0.22	0.74	0.29
E3SM-1-1-ECA	0.24	0.18	0.75	0.35
NESM3	0.17	0.44	0.77	0.15
***CNRM-ESM2-1 (A)**	0.22	0.69	0.55	0.10
BCC-CSM2-MR	0.26	0.10	1.00	0.23
ACCESS-ESM1-5	0.13	0.89	0.66	0.15
FGOALS-g3	0.48	0.59	0.69	0.13
***MIROC-ES2L (B)**	0.21	0.80	0.85	0.18
***MCM-UA-1-0 (C)**	0.38	0.72	0.80	0.16
***CAS-ESM2-0 (C)**	0.29	1.00	0.93	0.22
KIOST-ESM	0.66	0.21	0.86	1.00
GISS-E2-2-G	1.00	0.50	0.37	1.00
Multi-model mean	0.17	0.32	0.43	0.17
Multi-model spread	0.16	0.25	0.26	0.18

(b)

Summary statistics			
rank	RRPS	Storyline	Stats
1	0.06		Lower tail
2	0.07		”
3	0.12		”
4	0.14		”
5	0.14		”
6	0.15		”
7	0.16		”
8	0.16		”
9	0.16	B	”
10	0.17		”
11	0.17		”
12	0.17		IQR
13	0.17		”
14	0.19		”
15	0.20		”
16	0.20		”
17	0.20		”
18	0.21		”
19	0.22		”
20	0.22	C	”
21	0.23		”
22	0.23	C	”
23	0.23		”
24	0.23	B	”
25	0.25	B	”
26	0.25	A	”
27	0.26	D	”
28	0.26		”
29	0.28	C	”
30	0.29	D	”
31	0.30		”
32	0.31		”
33	0.31	A	”
34	0.32		”
35	0.32	C	”
36	0.34		”
37	0.34	A	”
38	0.35		Upper tail
39	0.36		”
40	0.38		”
41	0.38		”
42	0.39	A	”
43	0.40		”
44	0.46		”
45	0.47		”
46	0.51	B	”
47	0.51	C	”
48	0.61	C	Outliers
49	0.68		”
50	0.72		”
	0.27		
	0.15		

Figure 3 Overview of Arctic MJJASO NRMSE values, ranks, scores (RRPS), storylines, and quartile bins. Panel **a** is the same as Figure 2, but for the normalized RMSE (NRMSE) values. Panel **b** shows the ranks (column 1), scores (defined in section 3; column 2), the storylines for which the model is a candidate for (if any; column 3), and quartile bins (stats; column 4), indicating whether the model belongs to the lower tail, the IQR, the upper tail, or is an outlier. The IQR and outliers are highlighted in gray for readability. The models are sorted by the score, with the best model (lowest score) at the top.

where *fit* is the overall fit, *s* is the season, and *r* is the region.

Hence, the overall fit combines the relative present-day model performance with the representativeness of the candidate models into a single value. Values for the overall fit, based on the Arctic MJJASO scores for the candidate models and the Euclidean distances from Table 1, are illustrated in Figure 4 (blue curves) and given in Table 2.

For storyline A, CIESM (brown downward-pointing triangle) is the best-performing candidate model, while CNRM-CM6-1 (green filled circle) is slightly more representative of the storyline, having a smaller Euclidean distance (Figure 4a). The overall fit is best (smallest) for CIESM, meaning that even though CIESM is somewhat further away from the storyline point, it has a slightly better overall fit due to its present-day performance (i.e, score) for Arctic MJJASO. The overall fit for CIESM

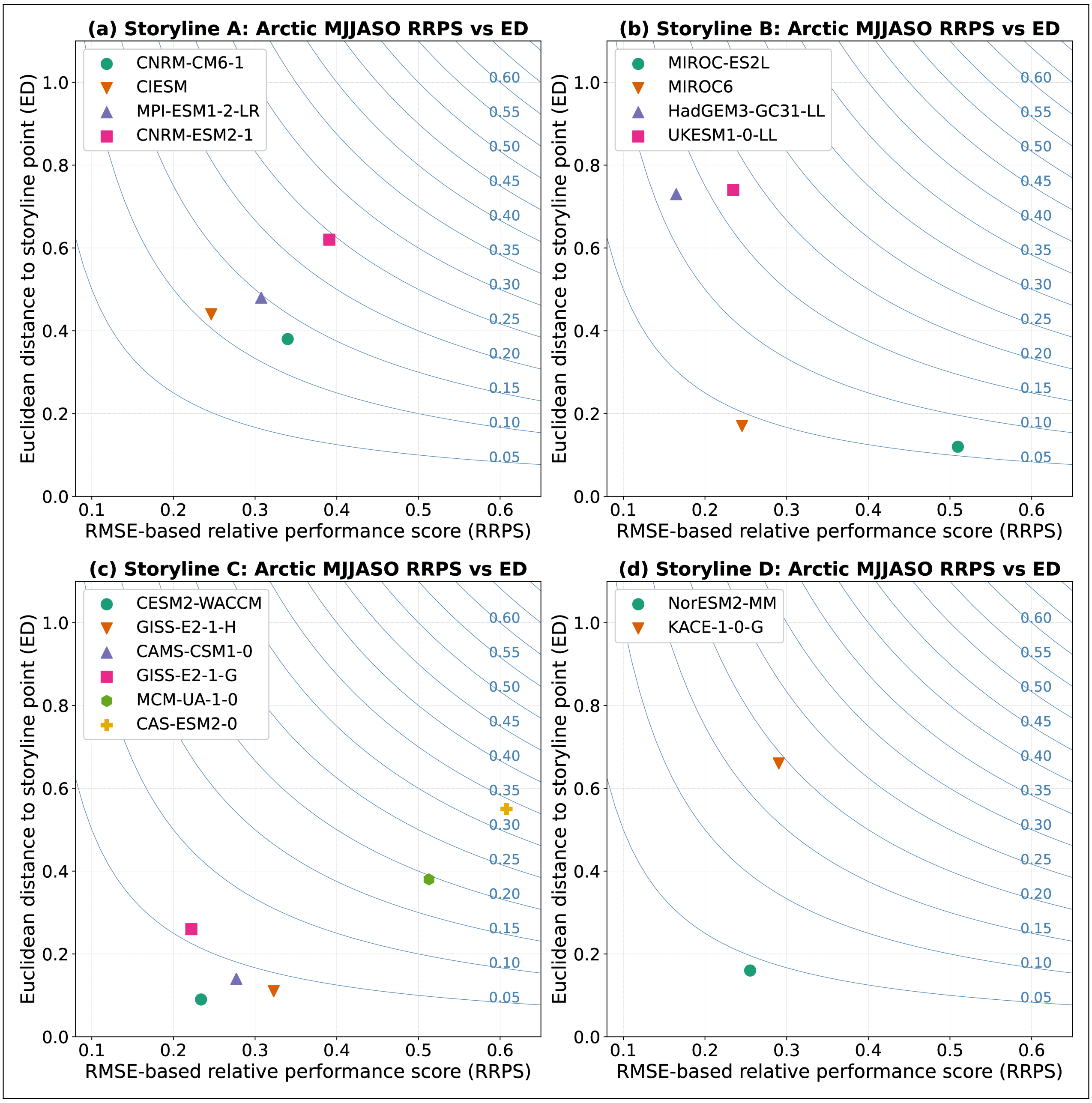


Figure 4 The Arctic MJJASO scores (RRPS) shown against the Euclidean distance (ED; Table 1) for candidate models (legends) for storyline A (panel **a**), B (panel **b**), C (panel **c**), and D (panel **d**). Also shown are isolines for the overall fit (blue curves with blue numbers), defined as the product of the score and the Euclidean distance (equation 5).

is nevertheless very similar to that for CNRM-CM6-1 and MPI-ESM1-2-LR (purple upward-pointing triangle). These three candidate models for storyline A are more or less clustered together with both scores and Euclidean distance that are relatively similar; we therefore argue that they should all be considered acceptable choices for storyline A. The last candidate model, CNRM-ESM2-1, is also relatively similar to the other three, but is further away from the storyline point and has a score exceeding the 75th percentile (Figure 3).

For storyline B (Figure 4b), MIROC-ES2L (green filled circle) and MIROC6 (brown downward-pointing triangle) stand out from the other candidate models due to their small Euclidean distances. While HadGEM3-GC31-LL (purple upward pointing triangle) and UKESM1-0-LL (pink square) have the smallest and second smallest scores, they are both much further away from the storyline point. MIROC6 stands out as the best candidate, having a small Euclidean distance and a relatively low score, and the best overall fit. MIROC-ES2L has the second-best fit, but it also has one of the largest scores of all the models considered (0.51) with high relative errors for both *pr* and *ua*850 (Figure 3). To choose a second model for storyline B, one will have to prioritize between having a small Euclidean distance and a low score. As MIROC-ES2L is in the far end of the upper tail, we argue that HadGEM3-GC31-LL is preferable in this case.

For storyline C (Figure 4c), it is clear what while CESM2-WACCM (green filled circle) has the smallest Euclidean distance and the second lowest score, yielding the best overall fit, results are very similar for GISS-E2-1-H (brown downward-pointing triangle), CAMS-CSM1-0 (purple upward-pointing triangle), and GISS-E2-1-G (pink square); hence, all these four models are good representatives for storyline C. The other two candidate models are less preferable, as they have larger Euclidean distances and substantially larger scores.

Storyline D has only two candidate models, which are more or less indistinguishable in terms of the Arctic MJJASO score alone; the models are both within the IQR, with the relative performance of NorESM2-MM being slightly better than for KACE-1-0-G (RRPS 0.26 and rank 27 vs. RRPS 0.29 and rank 30; Figure 3). While it is difficult to select one over the other based on the score alone, Figure 4d clearly shows NorESM2-MM (green filled circle) is much closer to the storyline point than KACE-1-0-G (brown downward-pointing triangle), resulting in the best overall fit. NorESM2-MM is therefore the preferred choice for storyline D.

In summary, based on a combination of (1) how well the models represent the present-day state of four key variables in the Arctic during MJJASO and (2) the models' proximity to the relevant storyline point in predictor space, we select CIESM, CNRM-CM6-1, and MPI-ESM1-2-LR to represent storyline A, MIROC6 to represent storyline B, CESM2-WACCM, GISS-E2-1-H, CAMS-CSM1-0, and GISS-E2-1-G to represent storyline C, and NorESM2-MM to represent storyline D. Storylines A and C both have multiple candidate models that are very similar, and we therefore recommend a set of models in these cases.

4.2.1 Performance for Arctic land and Arctic sea

Arctic climate change is associated with a wide range of impacts, some tied to processes over the land surface, such as wildfires and permafrost thaw (e.g., Chadburn *et al.*, 2017; Masrur *et al.*, 2018; McCarty *et al.*, 2021), and others to changes in marine areas, such as sea-ice loss (e.g., Gulev *et al.*, 2021; Screen and Simmonds, 2010). To assess how sensitive the model selection is to whether we base the score on NRMSE values for the whole Arctic, Arctic land, or Arctic sea, we compute the score and overall fit separately for these three regions (Table 2). Results show that the model with the best overall fit for the whole Arctic is generally also the model with the best fit for Arctic sea and Arctic land for all four storylines. For storyline A, CIESM has the best value for whole region and over the sea, while over land, CNRM-CM6-1 is marginally

STORYLINE	ARCTIC MJJASO RRPS			ARCTIC MJJASO OVERALL FIT		
	TOTAL	LAND	SEA	TOTAL	LAND	SEA
A	A2 (0.25)	A2 (0.24)	A3 (0.23)	A2 (0.11)	A1 (0.10)	A2 (0.10)
	A3 (0.31)	A1 (0.27)	A2 (0.24)	A1 (0.13)	A2 (0.10)	A3 (0.11)
	A1 (0.34)	A3 (0.32)	A1 (0.35)	A3 (0.15)	A3 (0.15)	A1 (0.13)
	A4 (0.39)	A4 (0.33)	A4 (0.38)	A4 (0.24)	A4 (0.20)	A4 (0.23)
B	B3 (0.16)	B3 (0.15)	B2 (0.16)	B2 (0.04)	B2 (0.05)	B2 (0.03)
	B4 (0.23)	B4 (0.21)	B3 (0.17)	B1 (0.06)	B1 (0.07)	B1 (0.04)
	B2 (0.25)	B2 (0.29)	B4 (0.24)	B3 (0.12)	B3 (0.11)	B3 (0.12)
	B1 (0.51)	B1 (0.55)	B1 (0.36)	B4 (0.17)	B4 (0.16)	B4 (0.18)
C	C4 (0.22)	C4 (0.21)	C1 (0.20)	C1 (0.02)	C1 (0.02)	C1 (0.02)
	C1 (0.23)	C1 (0.22)	C4 (0.23)	C2 (0.04)	C2 (0.03)	C2 (0.04)
	C3 (0.28)	C3 (0.25)	C3 (0.28)	C3 (0.04)	C3 (0.04)	C3 (0.04)
	C2 (0.32)	C2 (0.30)	C2 (0.33)	C4 (0.06)	C4 (0.05)	C4 (0.06)
	C5 (0.51)	C6 (0.52)	C5 (0.40)	C5 (0.19)	C5 (0.21)	C5 (0.15)
	C6 (0.61)	C5 (0.54)	C6 (0.62)	C6 (0.33)	C6 (0.29)	C6 (0.34)
D	D1 (0.26)	D1 (0.21)	D1 (0.27)	D1 (0.04)	D1 (0.03)	D1 (0.04)
	D2 (0.29)	D2 (0.26)	D2 (0.31)	D2 (0.19)	D2 (0.17)	D2 (0.20)

Table 2 Overview of MJJASO scores (RRPS) and overall fit for the whole (total) Arctic, Arctic land, and Arctic sea for the storyline candidate models. We use the model abbreviations defined in Table 1, repeated here for convenience: A1 (CNRM-CM6-1), A2 (CIESM), A3 (MPI-ESM1-2-LR), A4 (CNRM-ESM2-1), B1 (MIROC-ES2L), B2 (MIROC6), B3 (HadGEM3-GC31-LL), B4 (UKESM1-0-LL), C1 (CESM2-WACCM), C2 (GISS-E2-1-H), C3 (CAMS-CSM1-0), C4 (GISS-E2-1-G), C5 (MCM-UA-1-0), C6 (CAS-ESM2-0), D1 (NorESM2-MM), and D2 (KACE-1-0-G). For each storyline (column 1), the candidate-model abbreviations and their scores and overall fit for the whole Arctic, Arctic land, and Arctic sea are given in columns 2–4 and 5–7; the models are sorted by the score (columns 2–4) and overall fit (columns 5–7) with the best values on top.

better (however, at two decimal precision, the overall fit is the same).

While this shows that our model selection for Arctic MJJASO holds regardless of whether we focus on the whole Arctic, Arctic land, or Arctic ocean, it is clear that the scores display some sensitivity to which part of the Arctic they are computed for. Next, we examine the sensitivity of the score further, comparing results for Arctic MJJASO to results from other regions and seasons. This will allow us to further investigate how robust our selection is, that is, whether the model selection based on the Arctic MJJASO scores still holds for other regions and seasons, or whether other candidate models are preferable.

4.3 SEASONAL AND REGIONAL SENSITIVITY

Figure 5a compares the Arctic MJJASO scores (black dots in a) to the Arctic scores for the whole year and the four traditional three-month seasons (orange, blue, and red symbols), showing that the relative model performance can vary considerably throughout the year. The largest range (between the season with the worst and best score; numbers in the rightmost part of panel a) is found for GISS-E2-2-G, which has a range of 0.4 between the best-performing season (JJA; dark red square) and the worst (DJF; dark blue asterisk). There are no indications that the seasonal sensitivity increases linearly with the score, as large ranges are found for the models with low scores, small ranges are found for models with high scores, and vice versa.

Interestingly, the relative performance of the models varies more across regions than across seasons. Comparing the seasonal variability (for Arctic scores across seasons; Figure 5a) to the regional variability (for MJJASO scores across regions; Figure 5b) clearly shows that the regional variability is larger. This is not only

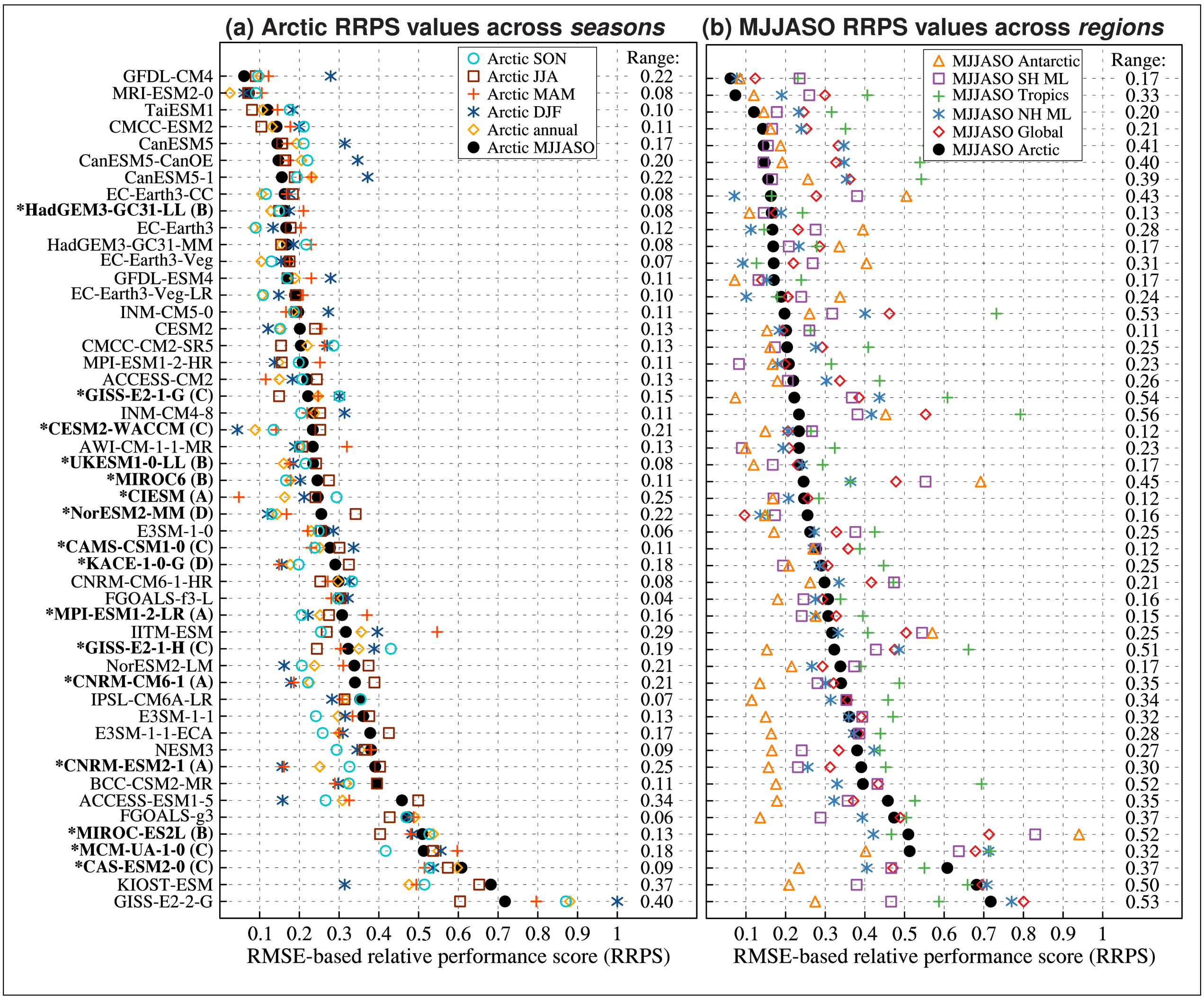


Figure 5 Arctic scores (RRPS) for different seasons **(a)** and MJJASO scores for different regions **(b)**. In (a), Arctic scores are shown for MJJASO (black dots), annual (orange diamonds), DJF (dark blue asterisks), MAM (light red plus signs), JJA (dark red squares), and SON (cyan open circles). In (b), MJJASO scores are shown for the Arctic (black dots), globe (red diamonds), NH mid-latitudes (NH ML; blue asterisk), tropics (green plus signs), SH mid-latitudes (SH ML; purple squares), and Antarctic (orange triangles). In both panels, the range between the smallest and largest scores for each model is given on the right side, and the models are sorted by the Arctic MJJASO scores (black dots). Models that are candidates for Arctic storylines are denoted as in Figure 2.

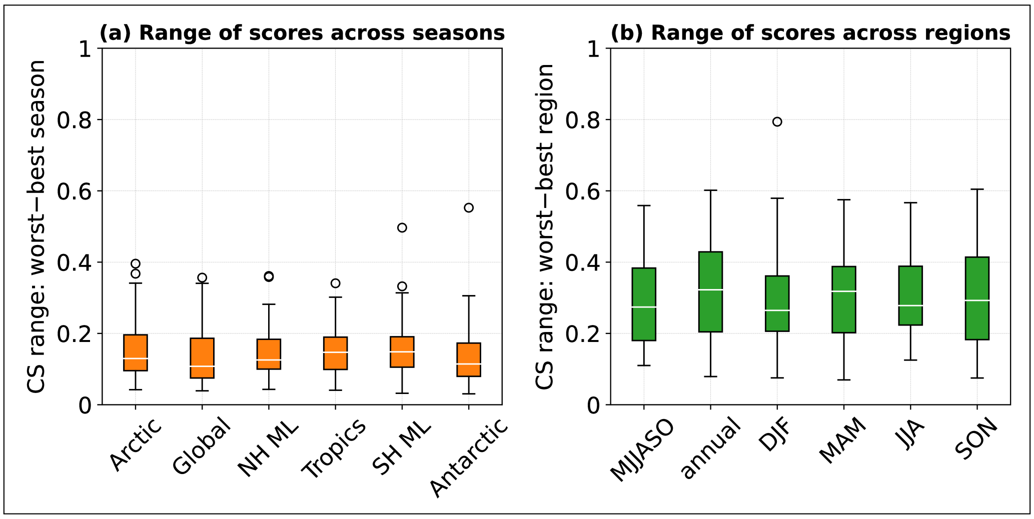


Figure 6 Distributions of the range of scores (RRPS) across seasons **(a)** and regions **(b)**. In (a), the ranges are defined as the difference between the *season* with the largest and smallest score for each model (as in Figure 5a), with the distributions based on the values from the 50 models shown separately for each *region* (orange boxes). In (b), the ranges are defined as the difference between the *region* with the largest and smallest score for each model (as in Figure 5b), with the distributions shown separately for each *season* (green boxes). The box and whiskers show the distribution for the 50 models with the boxes extending from the first to the third quartiles, the median shown as a white horizontal line, and the whiskers extending to the farthest data point or maximum 1.5 times the inter-quartile range. Scores that are more than 1.5 times the inter-quartile range from the box edge are defined as outliers and drawn as open circles.

evident when comparing the regional spread for a single season (Figure 5b) to the seasonal scores for a single region (Figure 5a), but is a general result found when considering the seasonal spread for all regions (Figure 6a) to the regional spread for all seasons at the same time (Figure 6b).

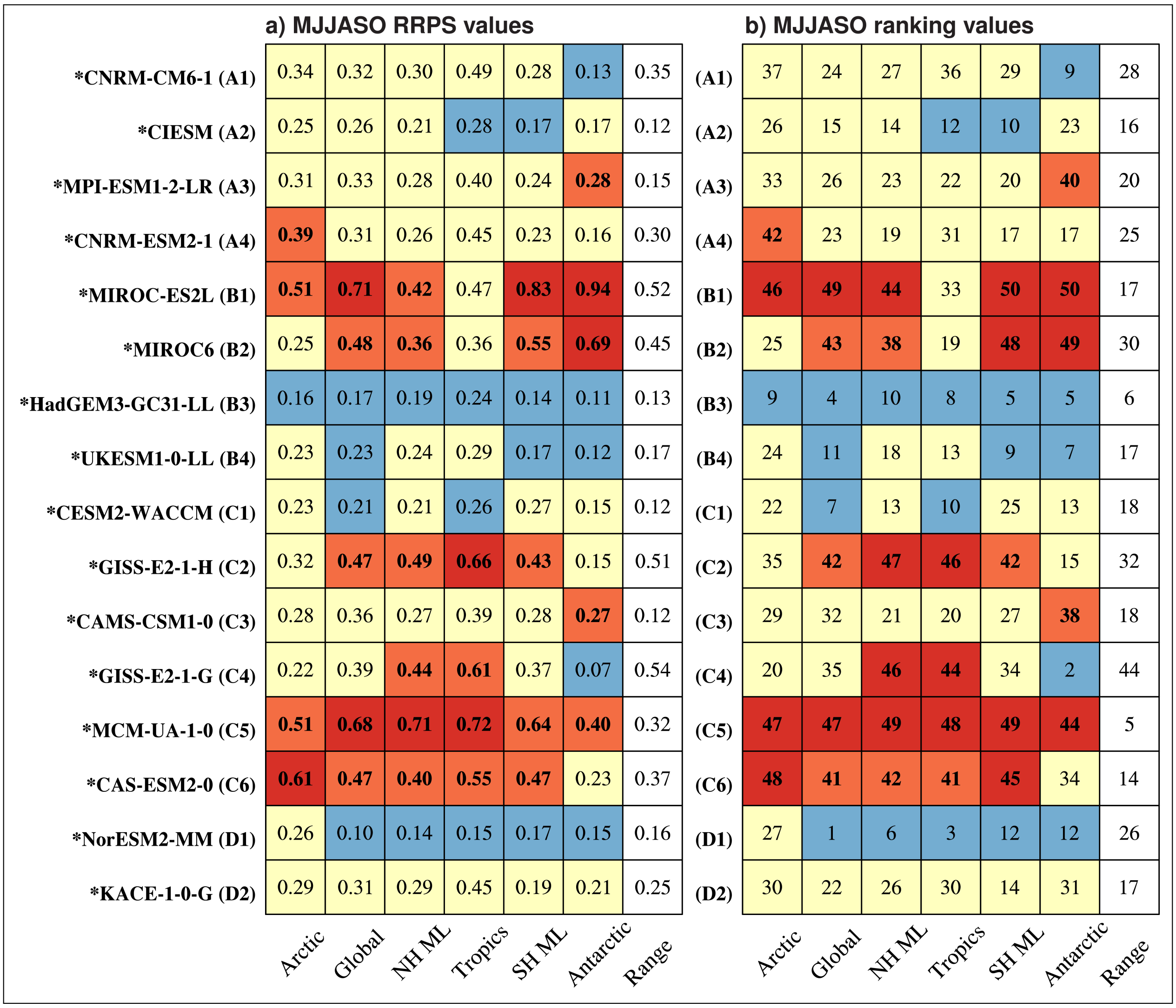
To further examine the regional sensitivity of the candidate models, we consider the MJJASO scores and ranks for these models for the different regions in Figure 7. While such an evaluation of the relative model performance is useful, it is important to keep in mind that both the scores and ranks are sensitive to the results for the other models, and that the sensitivity to the rest of the ensemble is stronger for the ranks. The score for an individual model is only affected by the subset of models that have the smallest or largest NRMSE value for one or more fields (equation 3). To find the ranking value of an individual model, on the other hand, the performance of the full set must be taken into account. Therefore, when comparing scores between regions for a single model, improved (lower) scores do not always translate to better (lower) ranking values, nor do worse scores always yield higher ranking values. For example, CNRM-ESM2-1 has a better score for the Antarctic (0.16) than for the SH mid-latitudes (0.23), but the ranking value is 17 in both cases. This happens because there are 16 better-performing models in both cases.

4.3.1 Storyline A

The models that were selected to represent storyline A based on the Arctic MJJASO scores, CNRM-CM6-1, CIESM, and MPI-ESM1-2-LR (Section 4.2), all have relatively consistent performance for the other regions (Figure 7a). Overall, the performance is best for CIESM, which has Arctic, global, NH mid-latitude, and Antarctic scores within the IQR (yellow cells) and tropical and SH mid-latitude scores below the 25th percentile (blue cells). CIESM is also the most consistent performer with a range of scores of 0.12. The second-best model is CNRM-CM6-1, which has scores within the IQR for every region except the Antarctic, where it is within the lower tail. MPI-ESM1-2-LR has scores within the IQR for all regions except the Antarctic, where it is ranked 40th with scores exceeding the 75th percentile (red cells). The last candidate model, CNRM-ESM2-1, is within the upper tail in the Arctic, but performs better for all other regions with scores within the IQR in all cases.

The overall fit for all seasons and regions (Figure 8a) shows that the CIESM (A2) is the overall best choice for storyline A, being either the best-performing model, or comparable to the best-performing model (e.g., in the Antarctic). To accentuate models with relatively good scores, we show markers for models with scores below the 75th percentile in strong colors (reds, blues, orange, and black), while markers for models whose scores exceed the 75th percentile are shown in light gray. The colors of the markers reveal that CIESM has no scores exceeding the 75th percentile.

While the overall fit for the candidate models for storyline A is best for CIESM, values for CNRM-CM6-1 and MPI-ESM1-2-LR are largely similar. CNRM-ESM2-1, however, tends to have slightly higher values, particularly in the tropics, and has Arctic scores exceeding the 75th percentile during MJJASO, JJA, and SON (gray markers).



a) MJJASO RRPS values

	Arctic	Global	NH ML	Tropics	SH ML	Antarctic	Range
*CNRM-CM6-1 (A1)	0.34	0.32	0.30	0.49	0.28	0.13	0.35
*CIESM (A2)	0.25	0.26	0.21	0.28	0.17	0.17	0.12
*MPI-ESM1-2-LR (A3)	0.31	0.33	0.28	0.40	0.24	**0.28**	0.15
*CNRM-ESM2-1 (A4)	**0.39**	0.31	0.26	0.45	0.23	0.16	0.30
*MIROC-ES2L (B1)	**0.51**	**0.71**	**0.42**	0.47	**0.83**	**0.94**	0.52
*MIROC6 (B2)	0.25	**0.48**	**0.36**	0.36	**0.55**	**0.69**	0.45
*HadGEM3-GC31-LL (B3)	0.16	0.17	0.19	0.24	0.14	0.11	0.13
*UKESM1-0-LL (B4)	0.23	0.23	0.24	0.29	0.17	0.12	0.17
*CESM2-WACCM (C1)	0.23	0.21	0.21	0.26	0.27	0.15	0.12
*GISS-E2-1-H (C2)	0.32	**0.47**	**0.49**	**0.66**	**0.43**	0.15	0.51
*CAMS-CSM1-0 (C3)	0.28	0.36	0.27	0.39	0.28	**0.27**	0.12
*GISS-E2-1-G (C4)	0.22	0.39	**0.44**	**0.61**	0.37	0.07	0.54
*MCM-UA-1-0 (C5)	**0.51**	**0.68**	**0.71**	**0.72**	**0.64**	**0.40**	0.32
*CAS-ESM2-0 (C6)	**0.61**	**0.47**	**0.40**	**0.55**	**0.47**	0.23	0.37
*NorESM2-MM (D1)	0.26	0.10	0.14	0.15	0.17	0.15	0.16
*KACE-1-0-G (D2)	0.29	0.31	0.29	0.45	0.19	0.21	0.25

b) MJJASO ranking values

	Arctic	Global	NH ML	Tropics	SH ML	Antarctic	Range
(A1)	37	24	27	36	29	9	28
(A2)	26	15	14	12	10	23	16
(A3)	33	26	23	22	20	**40**	20
(A4)	**42**	23	19	31	17	17	25
(B1)	**46**	**49**	**44**	33	**50**	**50**	17
(B2)	25	**43**	**38**	19	**48**	**49**	30
(B3)	9	4	10	8	5	5	6
(B4)	24	11	18	13	9	7	17
(C1)	22	7	13	10	25	13	18
(C2)	35	**42**	**47**	**46**	**42**	15	32
(C3)	29	32	21	20	27	**38**	18
(C4)	20	35	**46**	**44**	34	2	44
(C5)	**47**	**47**	**49**	**48**	**49**	**44**	5
(C6)	**48**	**41**	**42**	**41**	**45**	34	14
(D1)	27	1	6	3	12	12	26
(D2)	30	22	26	30	14	31	17

Figure 7 MJJASO scores (RRPS; **a**) and ranks **(b)** for the storyline candidate models for the Arctic (column 1), globe (column 2), NH mid-latitudes (NH ML; column 3), tropics (column 4), SH mid-latitudes (SH ML; column 5), and the Antarctic (column 6). Also shown is the range of values for each model (column 7), computed as the difference between the largest and smallest scores (a) and ranks (b) for each model. The colors indicate whether the values are within the lower tail (blue cells), within the IQR (yellow cells), within the upper tail (light red cells), or outlier values (dark red cells) based on percentiles computed separately for each region, using values from the full set of models. Bold values indicate that the scores (a) or ranks (b) exceed the 75th percentile. The model names follow the convention from Figure 2 and the sorting is as in Table 1. Note that the scores and ranks are relative to the full set of models.

4.3.2 Storyline B

MIROC6, the selected model for storyline B based on the Arctic MJJASO results (Section 4.2), consistently performs worse for the other regions, with MJJASO scores (Figure 7a) exceeding the 75th percentile for the globe, NH mid-latitudes, SH mid-latitudes, and the Antarctic, with the latter furthermore being an outlier. Similarly, the MJJASO scores for MIROC-ES2L exceed the 75th percentile for every region except the tropics, with the global, SH mid-latitude, and Antarctic scores being outliers. This suggests that when considering other regions than the Arctic, it is preferable to use one of the three other candidate models for storyline B, even though they all have larger distances to the storyline point (Table 1). Based on the MJJASO scores for the six regions (Figure 7a), HadGEM3-GC31-LL and UKESM1-0-LL are the best and second-best candidates. HadGEM3-GC31-LL consistently has scores in the lower tail for every region and also the smallest range of scores of the candidate models for storyline B. The ranks (Figure 7b) show that HadGEM3-GC31-LL is among the best 10 models for all regions. UKESM1-0-LL has scores in the IQR in the Arctic, NH mid-latitudes, and tropics and below the 25th percentile in the SH mid-latitudes, Antarctic, and globally, and is among the 24 best models in all cases.

Considering the overall fit for all seasons and regions (Figure 8b) reveals that MIROC6 (B2) generally has the best fit among the storyline B models or is comparable to MIROC-ES2L (B1), except in the Antarctic, where the overall fit of the four candidate models is relatively similar, with HadGEM3-CG31-LL (B3) and UKESM1-0-LL (B4) having somewhat better values for most seasons. However, the large scores seen for MJJASO outside the Arctic in Figure 7a are also found for other seasons.

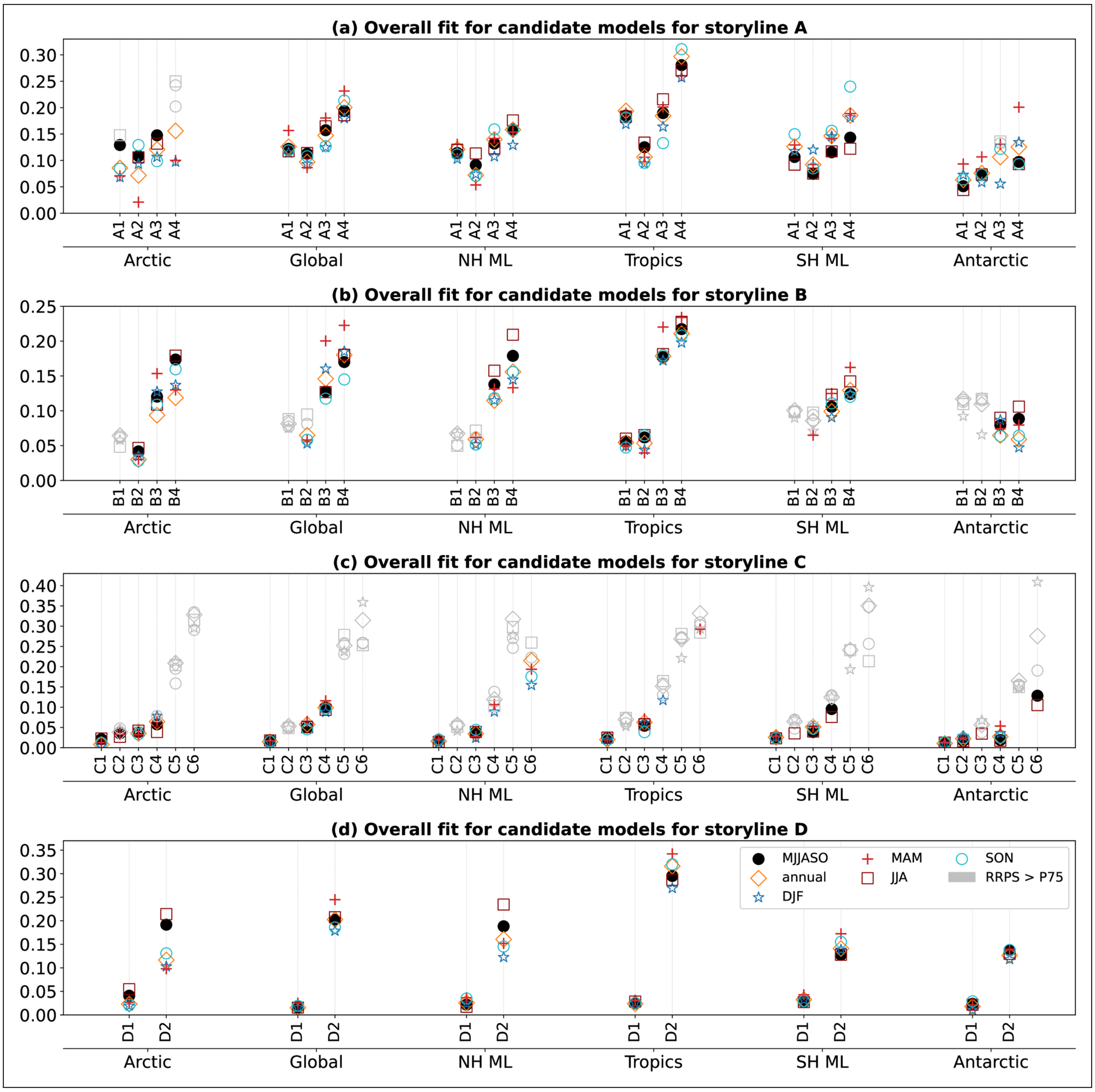


Figure 8 Overview of the overall fit (equation 5) for the candidate models for storylines A (panel **a**), B (panel **b**), C (panel **c**), and D (panel **d**) for all regions (Arctic, global, NH mid-latitudes (ML), tropics, SH ML, and Antarctic) and seasons (MJJASO, annual, DJF, MAM, JJA, and SON). We use the model abbreviations defined in Table 1, repeated here for convenience: A1 (CNRM-CM6-1), A2 (CIESM), A3 (MPI-ESM1-2-LR), A4 (CNRM-ESM2-1), B1 (MIROC-ES2L), B1 (MIROC6), B3 (HadGEM3-GC31-LL), B4 (UKESM1-0-LL), C1 (CESM2-WACCM), C2 (GISS-E2-1-H), C3 (CAMS-CSM1-0), C4 (GISS-E2-1-G), C5 (MCM-UA-1-0), C6 (CAS-ESM2-0), D1 (NorESM2-MM), and D2 (KACE-1-0-G). For each storyline, region, and candidate model, the overall fit for the whole year and the different seasons (legend in d) are shown in separate vertical stacks. To highlight the overall fit for models with relatively low scores, markers are shown in gray when the associated scores exceed the 75th percentile for the relevant region and season. Note that the y-axis varies between panels.

MIROC6 has scores above the 75th percentile (gray symbols) for one or more seasons for the global, NH and SH mid-latitude, and Antarctic regions. MIROC-ES2L has scores exceeding the 75th percentile for all regions and seasons except the tropics, where it is below the 75th percentile for all seasons.

While the MIROC models are both very close to the storyline point with small Euclidean distances, resulting in low (good) values for the overall fit, the relative model performance renders them less favorable for many regions. MIROC6 is the ideal choice for storyline B when considering the Arctic. For the tropics, MIROC6 and MIROC-ES2L are both good choices. For the other regions, HadGEM-GC-31-LL and UKESM1-0-LL are the preferred choices; these two models have scores below the 75th percentile for all seasons and regions. While HadGEM-GC-31-LL tends to have slightly better overall performance, the values for UKESM1-0-LL are generally similar.

4.3.3 Storyline C

For storyline C, CESM2-WACCM, GISS-E2-1-H, CAMS-CSM1-0, and GISS-E2-1-G were selected based on the

Arctic MJJASO scores and overall fit. Considering MJJASO scores for all regions (Figure 7a), CESM2-WACCM is the best performer overall with scores below the 25th percentile in the tropics and globally and within the IQR in the other regions. CESM2-WACCM is also the most consistent performer of all the candidate models for storyline C, with a range of 0.12. The other candidate models for storyline C all exceed the 75th percentile for at least one region.

The overall fit (Figure 8c) is consistently better for the four models that were selected to represent storyline C based on the Arctic MJJASO results (C1–C4) compared to the other models (C5–C6). However, CESM2-WACCM (C1) is the only model for which the scores stay below the 75th percentile across all regions and seasons. CESM2-WACCM therefore stands out as the best choice for storyline C based on the overall fit and scores for all seasons and regions.

4.3.4 Storyline D

Storyline D stands out from the others in that it only has two candidate models, NorESM2-MM and KACE-1-0-G (Table 1). While the Arctic MJJASO scores are similar for these two models, NorESM2-MM has the best overall fit and was selected to represent storyline D in Section 4.2. Considering the MJJASO scores for all regions (Figure 7a) shows that while the scores are similar for many regions, including the Arctic, NorESM2-MM consistently has better (lower) scores, particularly in the NH mid-latitudes, tropics, Antarctic, and globally. The ranks (b) show that NorESM2-MM is the best-ranking model in terms of the global scores, the third-best ranking model in the tropics, and the sixth-best ranking model in the NH mid-latitudes. Based on MJJASO scores for the globe, NH mid-latitudes, tropics, and Antarctic, NorESM2-MM is the preferred model for storyline D with scores below the 25th percentile for all regions except the Arctic. For the Arctic and SH mid-latitudes, while NorESM2-MM has slightly better scores than KACE-1-0-G, the models are relatively similar in terms of their relative performance.

When considering the overall fit for all seasons and regions (Figure 8d), NorESM2-MM (D1) clearly stands out as the best choice in all cases. This is in line with the NorESM2-MM being closer to the storyline point and hence having a smaller Euclidean distance than KACE-1-0-G (Table 1) and with the scores generally being similar for the two models or better for the NorESM2-MM. Both models consistently have scores below the 75th percentile (i.e., no gray symbols).

5 SUMMARY AND DISCUSSION

This study presents a novel framework for evaluating the historical representation of climate models, based on NRMSE (normalized RMSE) for multiple variables of particular interest, yielding a single number, or score, per model: the RMSE-based relative performance score (RRPS). The novelty lies in the way the normalization is performed, forcing all variables to vary within the exact same range (0 to 1), thus exerting a comparable influence on the score. Some differences between variables, however, remain, as the NRMSE distributions themselves are different. The score is an easily understandable and implementable way of evaluating relative model performance for variables, regions, and seasons of particular interest. It is inherently flexible in that different variables and statistics that underpin a specific study can be used, as long as they are normalized in the same way. The results highlight models with large errors relative to the multi-model ensemble, and facilitate identifying models that perform less favorably relative to other models within the ensemble, thus providing a quantifiable and objective approach to narrowing the selection of models in studies that cannot use all models, even though all models may perform acceptably well.

We demonstrate the benefits and limitations of the score through the selection of specific CMIP6 models that represent previously defined storylines of Arctic climate change (Levine *et al.*, 2024): *weak* Arctic amplification and *strong* Barents-Kara Sea warming (storyline A), *strong* Arctic amplification and *strong* Barents-Kara Sea warming (storyline B), *weak* Arctic amplification and *weak* Barents-Kara Sea warming (storyline C), and *strong* Arctic amplification and *weak* Barents-Kara Sea warming (storyline D). We achieve this through a three-step process:

1. We identify a set of CMIP6 models that are close to the storylines in terms of their future changes in Arctic amplification and Barents-Kara Sea warming and estimate this closeness in terms of the Euclidean distance between the storyline point and the models in predictor space (Table 1).
2. We use the score to evaluate the models' present-day performance, considering data from the historical experiments from 50 CMIP6 models and reference data from ERA5 and GPCP for four key variables used in Levine *et al*. (2024): *tas*, *pr*, *ta*850, and *ua*850 for 1985–2014.
3. We combine the Euclidean distance and the scores to produce an estimate of the overall fit of each model (equation 5) and use this as a basis for the final model selection.

We focus on the Arctic during the extended summer season (MJJASO), in line with the region and season used in Levine *et al*. (2024), and find CIESM, CNRM-CM6-1, and MPI-ESM1-2-LR to be the best models for representing storyline A, MIROC6 for storyline B, CESM2-WACCM, GISS-E2-1-H, CAMS-CSM1-0, and GISS-E2-1-G for storyline C, and NorESM2-MM for storyline D.

Assessing the robustness of our results, we also consider scores for a comprehensive set of regions and seasons. The score exhibits both seasonal and regional sensitivity, with the regional sensitivity being larger than the seasonal sensitivity. For storyline A, the selection based on Arctic MJJASO results holds across regions and seasons, albeit with the CIESM standing out as a somewhat better choice than CNRM-CM6-1 and MPI-ESM1-2-LR. For storyline B, we find the best choices to be MIROC6 for the Arctic, MIROC6 and MIROC-ES2L for the tropics, and HadGEM-GC-21-LL and UKESM1-0-LL for the other regions (global, NH and SH extratropics, and Antarctic). For storylines C and D, we find CESM2-WACCM and NorESM2-MM, respectively, to be the best choices.

For any application of the score, the variables must be carefully selected to capture the most important aspects of the research topic. The set considered here is tailored for the Arctic storylines in Levine *et al*. (2024); many other properties of the modeled climate system could have been evaluated, and the analysis presented here is not an exhaustive investigation of the general performance of the models. While we wanted the variables considered to have a comparable influence on the score, it can be desirable to amplify or lessen the influence of some variables, for example, based on skill, observational uncertainty, or co-variability between variable pairs, for other applications. While beyond the scope of this study, such adjustments can be incorporated in the normalization itself or as weights assigned to the NRMSE values before averaging. The weights must, however, be tailored for the season and region of interest, as co-variability between variables can have pronounced seasonal and regional sensitivity.

In some cases, it can be deemed necessary to exclude outlier models from the ensemble. This can have a considerable impact on the score considered here, as the largest (and smallest) RMSE values are used in the normalization (equation 3), hence affecting the scoring values of all models. The score is, on the other hand, relatively insensitive to the inclusion of multiple models with similar errors, in contrast to ranking-based methods. Here, we opt for considering the full multi-model ensemble to ensure that the same set of models is examined for all regions and seasons and that all relevant candidate models are always included. As mentioned above, model performance can vary substantially from region to region and season to season, resulting in different outlier models for different regions and seasons.

We use the score and Euclidean distance to identify a subset of the candidate models that we consider to be more suitable than the others. However, for some applications of the storylines, it can be necessary to introduce other or additional criteria that can modify the outcome. An example of this is the availability of high-frequency data for downscaling.

The purpose of the score is not to identify the best-performing models, but to identify models whose performance deviates from the rest of the multi-model ensemble for the variables, regions, and seasons of interest. The score presented here provides a relative measure of quality compared to other models, and hence does not say anything about how good or bad a model is in absolute terms. In an ensemble of excellent performing models, the worst can still be well-suited, and in a group of models that perform terribly, even the best may represent the climate so poorly that it should be used with caution.

APPENDIX A: CMIP6 MODELS

In this study, we use data from the CMIP6 historical experiments of the following 50 models (the realization/variant label and data citation are provided in parenthesis): ACCESS-CM2 (r1i1p1f1; Dix *et al.*, 2019); ACCESS-ESM1-5 (r1i1p1f1; Ziehn *et al.*, 2019); AWI-CM-1-1-MR (r1i1p1f1; Semmler *et al.*, 2018); BCC-CSM2-MR (r1i1p1f1; Wu *et al.*, 2018); CAMS-CSM1-0 (r1i1p1f1; Rong, 2019); CAS-ESM2-0 (r3i1p1f1; Chai, 2020); CESM2 (r1i1p1f1; Danabasoglu, 2019a); CESM2-WACCM (r1i1p1f1; Danabasoglu, 2019b); CIESM (r1i1p1f1; Huang, 2019); CMCC-CM2-SR5 (r1i1p1f1; Lovato and Peano, 2020); CMCC-ESM2 (r1i1p1f1; Lovato *et al.*, 2021); CNRM-CM6-1 (r1i1p1f2; Voldoire, 2018); CNRM-CM6-1-HR (r1i1p1f2; Voldoire, 2019); CNRM-ESM2-1 (r1i1p1f2; Seferian, 2018); CanESM5 (r1i1p1f1; Swart *et al.*, 2019b); CanESM5-1 (r1i1p1f1; Swart *et al.*, 2019c); CanESM5-CanOE (r1i1p2f1; Swart *et al.*, 2019a); E3SM-1-0 (r1i1p1f1; Stevenson *et al.*, 2023); E3SM-1-1 (r1i1p1f1; Bader *et al.*, 2019); E3SM-1-1-ECA (r1i1p1f1; Bader *et al.*, 2020); EC-Earth3 (r1i1p1f1; EC-Earth Consortium (EC-Earth), 2019a); EC-Earth3-CC (r1i1p1f1; EC-Earth Consortium (EC-Earth), 2021); EC-Earth3-Veg (r1i1p1f1; EC-Earth Consortium (EC-Earth), 2019b); EC-Earth3-Veg-L (r1i1p1f1; EC-Earth Consortium (EC-Earth), 2020); FGOALS-f3-L (r1i1p1f1; Yu, 2019); FGOALS-g3 (r1i1p1f1; Li, 2019); GFDL-CM4 (r1i1p1f1; Guo *et al.*, 2018); GFDL-ESM4 (r1i1p1f1; Krasting *et al.*, 2018); GISS-E2-1-G (r3i1p5f1; NASA Goddard Institute for Space Studies (NASA/GISS), 2018); GISS-E2-1-H (r5i1p1f2; NASA Goddard Institute for Space Studies (NASA/GISS), 2019b); GISS-E2-2-G (r1i1p3f1; NASA Goddard Institute for Space Studies (NASA/GISS), 2019a); HadGEM3-GC31-LL (r3i1p1f3; Ridley *et al.*, 2019a); HadGEM3-GC31-MM (r1i1p1f3; Ridley *et al.*, 2019b); IITM-ESM (r1i1p1f1; Choudhury *et al.*, 2019); INM-CM4-8 (r1i1p1f1; Volodin *et al.*, 2019a); INM-CM5-0 (r1i1p1f1; Volodin *et al.*, 2019b); IPSL-CM6A-LR (r1i1p1f1; Boucher *et al.*, 2018); KACE-1-0-G (r1i1p1f1; Byun *et al.*, 2019); KIOST-ESM (r1i1p1f1; Kim *et al.*, 2019); MCM-UA-1-0 (r1i1p1f2; Stouffer, 2019); MIROC-ES2L (r8i1p1f2; Hajima *et al.*, 2019); MIROC6 (r23i1p1f1; Tatebe and Watanabe, 2018); MRI-ESM1-2-HR (r1i1p1f1; Jungclaus *et al.*, 2019); MPI-ESM1-2-LR (r17i1p1f1; Wieners *et al.*, 2019); MRI-ESM2-0 (r1i1p1f1; Yukimoto *et al.*, 2019); NESM3 (r1i1p1f1; Cao and Wang, 2019); NorESM2-LM (r1i1p1f1; Seland *et al.*, 2019); NorESM2-MM (r1i1p1f1; Bentsen *et al.*, 2019); TaiESM1 (r1i1p1f1; Lee and Liang, 2020); UKESM1-0-LL (r8i1p1f2; Tang *et al.*, 2019).

Note that we use the first realization (r1), except for the candidate models for Arctic storylines, where we use the realization that is closest to the storyline point in predictor space (Section 4.1 and Table 1).

DATA ACCESSIBILITY STATEMENT

An overview of the CMIP6 data is provided in Appendix A. The CMIP6 data is freely available through the Earth System Grid Federation, see for instance https://esgf.github.io/nodes.html. Data from ERA5 can be retrieved through the Copernicus Climate Data Store (https://cds.climate.copernicus.eu) and the GPCP data from the National Oceanic and Atmospheric Administration Physical Science Laboratory website (https://psl.noaa.gov). Data citations are provided in Section 2 and Appendix A.

Regridding the CMIP6 data to a 1 × 1 common grid and computing monthly climatologies (steps 1 and 2 in Section 3) was carried out in ESMValTool version 2.10 (Eyring *et al.*, 2016b; Righi *et al.*, 2020). Subsequently, the remaining steps (3–7 in Section 3) were carried out in NCL.

ACKNOWLEDGEMENTS

We acknowledge the World Climate Research Programme for coordinating and promoting CMIP6 through its Working Group on Coupled Modelling, the individual climate modeling groups for producing the model output and making it available, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies that support CMIP6 and ESGF.

FUNDING INFORMATION

The authors were supported by PolarRES (grant number 101003590), a project funded under the European Union's Horizon 2020 research and innovation program. In addition, LSG received funding from the Research Council of Norway project KeyCLIM (project number 295046) and OAL also received funding from the NeIC NICEST2 project, which is supported by the Nordic e-Infrastructure Collaboration (NeIC), part of NordForsk. Storage resources were provided by Sigma2—the National Infrastructure for High Performance Computing and Data Storage in Norway (projects NS8002K, NS9252K, and NS9560K).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

PAM secured the primary funding for the work. All authors contributed to developing the idea behind the study. LSG coordinated the study, wrote the code for computing the RRPS, and carried out the analysis for Sections 4.2 and 4.3. XL carried out the analysis for Section 4.1. LSG, OAL, and KMP wrote the first draft. All authors provided feedback and suggestions throughout the process and reviewed and contributed to several versions of the manuscript.

AUTHOR AFFILIATIONS

Lise Seland Graff orcid.org/0000-0003-3217-6329
Norwegian Meteorological Institute, Oslo, Norway

Oskar A. Landgren orcid.org/0000-0002-6264-8502
Norwegian Meteorological Institute, Oslo, Norway

Kajsa M. Parding orcid.org/0000-0001-6840-7243
Norwegian Meteorological Institute, Oslo, Norway

Xavier Levine orcid.org/0000-0003-4970-7026
NORCE Research AS, Bjerknes Centre for Climate Research, Bergen, Norway

Ryan S. Williams orcid.org/0000-0002-3185-3909
British Antarctic Survey, Cambridge, United Kingdom

Priscilla A. Mooney orcid.org/0000-0001-5921-3105
NORCE Research AS, Bjerknes Centre for Climate Research, Bergen, Norway

REFERENCES

Adler, R., Huffman, G.J., Chang, A., Ferraro, R., Xie, P.P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P. and **Nelkin, E.** (2003) The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present). *Journal of Hydrometeorology*, 4(6): 1147–1167. DOI: https://doi.org/10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2

Adler, R., Wang, J.J., Sapiano, M., Huffman, G., Chiu, L., Xie, P.P., Ferraro, R., Schneider, U., Becker, A., Bolvin, D., Nelkin, E., Gu, G. and **Program, N.C.** (2016) *Global Precipitation Climatology Project (GPCP) Climate Data Record (CDR), Version 2.3 (Monthly)*. National Centers for Environmental Information. DOI: https://doi.org/10.7289/V56971M6 (Accessed: 2022-04-04).

Ashfaq, M., Rastogi, D., Kitson, J., Abid, M.A. and **Kao, S.C.** (2022) Evaluation of CMIP6 GCMs over the CONUS for downscaling studies. *Journal of Geophysical Research: Atmospheres*, 127(21): e2022JD036659. DOI: https://doi.org/10.1029/2022JD036659

Bader, D.C., Leung, R., Taylor, M. and **McCoy, R.B.** (2019) *E3SM-Project E3SM1.1 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.11485

Bader, D.C., Leung, R., Taylor, M. and **McCoy, R.B.** (2020) *E3SM-Project E3SM1.1ECA model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.11486

Benestad, R.E., Mezghani, A., Lutz, J., Dobler, A., Parding, K.M. and **Landgren, O.A.** (2023) Various ways of using Empirical Orthogonal Functions for Climate Model evaluation. *EGUsphere*, 2023: 1–25. DOI: https://doi.org/10.5194/gmd-16-2899-2023

Bentsen, M., Oliviè, D.J.L., Seland, Ø., Toniazzo, T., Gjermundsen, A., Graff, L.S., Debernard, J.B., Gupta, A.K., He, Y., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K.S., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I.H.H., Landgren, O.A., Liakka, J., Moseid, K.O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T. and **Schulz, M.** (2019) *NCC NorESM2-MM model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.8040

Boucher, O., Denvil, S., Levavasseur, G., Cozic, A., Caubel, A., Foujols, M.A., Meurdesoif, Y., Cadule, P., Devilliers, M., Ghattas, J., Lebas, N., Lurton, T., Mellul, L., Musat, I., Mignot, J. and **Cheruy, F.** (2018) *IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.5195

Brands, S. (2022) A circulation-based performance atlas of the CMIP5 and 6 models for regional climate studies in the Northern Hemisphere mid-to-high latitudes. *Geoscientific Model Development*, 15(4): 1375–1411. DOI: https://doi.org/10.5194/gmd-15-1375-2022

Byun, Y.H., Lim, Y.J., Sung, H.M., Kim, J., Sun, M. and **Kim, B.H.** (2019) *NIMS-KMA KACE1.0-G model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.8378

Cao, J. and **Wang, B.** (2019) *NUIST NESMv3 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.8769

Chadburn, S.E., Burke, E.J. Cox, P.M., Friedlingstein, P., Hugelius, G. and **Westermann, S.** (2017) An observation-based constraint on permafrost loss as a function of global warming. *Nature Climate Change*, 7: 340–344. DOI: https://doi.org/10.1038/nclimate3262

Chai, Z. (2020) *CAS CAS-ESM1.0 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.3353

Choudhury, A.D., Raghavan, K., Gopinathan, P.A., Narayanasetti, S., Singh, M., Panickal, S. and **Modi, A.** (2019) *CCCR-IITM IITM-ESM model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.3708

Danabasoglu, G. (2019a) *NCAR CESM2 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.7627

Danabasoglu, G. (2019b) *NCAR CESM2-WACCM model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.10071

Dix, M., Bi, D., Dobrohotoff, P., Fiedler, R., Harman, I., Law, R., Mackallah, C., Marsland, S., O'Farrell, S., Rashid, H., Srbinovsky, J., Sullivan, A., Trenham, C., Vohralik, P., Watterson, I., Williams, G., Woodhouse, M., Bodman, R., Dias, F.B., Domingues, C.M., Hannah, N., Heerdegen, A., Savita, A., Wales, S., Allen, C., Druken, K., Evans, B., Richards, C., Ridzwan, S.M., Roberts, D., Smillie, J., Snow, K., Ward, M. and **Yang, R.** (2019) *CSIRO-ARCCSS ACCESS-CM2 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.4271

EC-Earth Consortium (EC-Earth) (2019a) *EC-Earth-Consortium EC-Earth3 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.4700

EC-Earth Consortium (EC-Earth) (2019b) *EC-Earth-Consortium EC-Earth3-Veg model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.4706

EC-Earth Consortium (EC-Earth) (2020) *EC-Earth-Consortium EC-Earth3-Veg-LR model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.4707

EC-Earth Consortium (EC-Earth) (2021) *EC-Earth-Consortium EC-Earth-3-CC model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.4702

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J. and **Taylor, K.E.** (2016a) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5): 1937–1958. DOI: https://doi.org/10.5194/gmd-9-1937-2016

Eyring, V., Cox, P.M., Flato, G.M., Gleckler, P.J., Abramowitz, G., Caldwell, P., Collins, W.D., Gier, B.K., Hall, A.D., Hoffman, F.M., Hurtt, G.C., Jahn, A., Jones, C.D., Klein, S.A., Krasting, J.P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G.A., Pendergrass, A.G., Pincus, R., Ruane, A.C., Russell, J.L., Sanderson, B.M., Santer, B.D., Sherwood, S.C., Simpson, I.R., Stouffer, R.J. and **Williamson, M.S.** (2019) Taking climate model evaluation to the next level. *Nature Climate Change*, 9(2): 102–110. DOI: https://doi.org/10.1038/s41558-018-0355-y

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., Anav, A., Andrews, O., Cionni, I., Davin, E.L., Deser, C., Ehbrecht, C., Friedlingstein, P., Gleckler, P., Gottschaldt,

K.D., Hagemann, S., Juckes, M., Kindermann, S., Krasting, J., Kunert, D., Levine, R., Loew, A., Mäkelä, J., Martin, G., Mason, E., Phillips, A.S., Read, S., Rio, C., Roehrig, R., Senftleben, D., Sterl, A., van Ulft, L., Walton, J., Wang, S. and **Williams, K.** (2016b) ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geoscientific Model Development*, 9(5): 1747–1802. DOI: https://doi.org/10.5194/gmd-9-1747-2016

Gleckler, P.J., Taylor, K.E. and **Doutriaux, C.** (2008) Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, 113(D6). DOI: https://doi.org/10.1029/2007JD008972

Gulev, S.K., Thorne, P., Ahn, J., Dentener, F.J., Domingues, C.M., Gerland, S., Gong, D., Kaufman, D., Nnamchi, H., Quaas, J., Rivera, J.A., Sathyendranath, S., Smith, S.L., Trewin, B., von Schuckmann, K. and **Vose, R.** (2021) *Changing State of the Climate System*. Cambridge University Press. pp. 287–422.

Guo, H., John, J.G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A., Rand, K., Zadeh, N.T., Balaji, V., Durachta, J., Dupuis, C., Menzel, R., Robinson, T., Underwood, S., Vahlenkamp, H., Bushuk, M., Dunne, K.A., Dussin, R., Gauthier, P.P., Ginoux, P., Griffies, S.M., Hallberg, R., Harrison, M., Hurlin, W., Lin, P., Malyshev, S., Naik, V., Paulot, F., Paynter, D.J., Ploshay, J., Reichl, B.G., Schwarzkopf, D.M., Seman, C.J., Shao, A., Silvers, L., Wyman, B., Yan, X., Zeng, Y., Adcroft, A., Dunne, J.P., Held, I.M., Krasting, J.P., Horowitz, L.W., Milly, P., Shevliakova, E., Winton, M., Zhao, M. and **Zhang, R.** (2018) *NOAA-GFDL GFDL-CM4 model output historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.8594

Hajima, T., Abe, M., Arakawa, O., Suzuki, T., Komuro, Y., Ogura, T., Ogochi, K., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M.A., Ohgaito, R., Ito, A., Yamazaki, D., Ito, A., Takata, K., Watanabe, S., Kawamiya, M. and **Tachiiri, K.** (2019) *MIROC MIROC-ES2L model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.5602

Hannachi, A. (2021) *Patterns Identification and Data Mining in Weather and Climate*. Cham: Springer. DOI: https://doi.org/10.1007/978-3-030-67073-3

Hannachi, A., Finke, K. and **Nickolay, T.** (2022) Common EOFs: A tool for multi-model comparison and evaluation. *Climate Dynamics*, 60: 1689–1703. DOI: https://doi.org/10.1007/s00382-022-06409-8

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D. and **Thépaut, J.N.** (2019a) ERA5 monthly averaged data on pressure levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: https://doi.org/10.24381/cds.6860a573 (Accessed: 2022-03-29).

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D. and **Thépaut, J.N.** (2019b) ERA5 monthly averaged data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: https://doi.org/10.24381/cds.f17050d7 (Accessed: 2022-04-04).

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and **Thépaut, J.N.** (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049. DOI: https://doi.org/10.1002/qj.3803

Hu, Z., Chen, D., Chen, X., Zhou, Q., Peng, Y., Li, J. and **Sang, Y.** (2022) CCHZ-DISO: A timely new assessment system for data quality or model performance from Da Dao Zhi Jian. *Geophysical Research Letters*, 49(23): e2022GL100681. DOI: https://doi.org/10.1029/2022GL100681

Huang, W. (2019) *THU CIESM model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.8843

Jungclaus, J., Bittner, M., Wieners, K.H., Wachsmann, F., Schupfner, M., Legutke, S., Giorgetta, M., Reick, C., Gayler, V., Haak, H., de Vrese, P., Raddatz, T., Esch, M., Mauritsen, T., von Storch, J.S., Behrens, J., Brovkin, V., Claussen, M., Crueger, T., Fast, I., Fiedler, S., Hagemann, S., Hohenegger, C., Jahns, T., Kloster, S., Kinne, S., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Müller, W., Nabel, J., Notz, D., Peters-von Gehlen, K., Pincus, R., Pohlmann, H., Pongratz, J., Rast, S., Schmidt, H., Schnur, R., Schulzweida, U., Six, K., Stevens, B., Voigt, A. and **Roeckner, E.** (2019) *MPI-M MPI-ESM1.2-HR model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.6594

Karpechko, A.Y., Gillett, N.P., Hassler, B., Rosenlof, K.H. and **Rozanov, E.** (2010) Quantitative assessment of Southern Hemisphere ozone in chemistry-climate model simulations. *Atmospheric Chemistry and Physics*, 10(3): 1385–1400. DOI: https://doi.org/10.5194/acp-10-1385-2010

Kim, Y., Noh, Y., Kim, D., Lee, M.I., Lee, H.J., Kim, S.Y. and **Kim, D.** (2019) *KIOST KIOST-ESM model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.5296

Krasting, J.P., John, J.G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A., Rand, K., Zadeh, N.T., Balaji, V., Durachta, J., Dupuis, C., Menzel, R., Robinson, T., Underwood, S., Vahlenkamp, H., Dunne, K.A., Gauthier, P.P., Ginoux, P., Griffies, S.M., Hallberg, R., Harrison, M., Hurlin, W., Malyshev, S., Naik, V., Paulot, F., Paynter, D.J., Ploshay, J., Reichl, B.G., Schwarzkopf, D.M., Seman, C.J.,

Silvers, L., Wyman, B., Zeng, Y., Adcroft, A., Dunne, J.P., Dussin, R., Guo, H., He, J., Held, I.M., Horowitz, L.W., Lin, P., Milly, P., Shevliakova, E., Stock, C., Winton, M., Wittenberg, A.T., Xie, Y. and **Zhao, M.** (2018) *NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.8597

Lee, W.L. and **Liang, H.C.** (2020) *AS-RCEC TaiESM1.0 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.9755

Levine, X.J., Williams, R.S., Marshall, G., Orr, A., Seland Graff, L., Handorf, D., Karpechko, A., Köhler, R., Wijngaard, R.R., Johnston, N., Lee, H., Nieradzik, L. and **Mooney, P.A.** (2024) Storylines of summer Arctic climate change constrained by Barents–Kara seas and Arctic tropospheric warming for climate risk assessment. *Earth System Dynamics*, 15(4): 1161–1177. DOI: https://doi.org/10.5194/esd-15-1161-2024

Li, L. (2019) *CAS FGOALS-g3 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.3356

Lovato, T. and **Peano, D.** (2020) *CMCC CMCC-CM2-SR5 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.3825

Lovato, T., Peano, D. and **Butenschön, M.** (2021) *CMCC CMCC-ESM2 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.13195

Masrur, A., Petrov, A.N. and **DeGroote, J.** (2018) Circumpolar spatio-temporal patterns and contributing climatic factors of wildfire activity in the Arctic tundra from 2001–2015. *Environmental Research Letters*, 13(1): 014019. DOI: https://doi.org/10.1088/1748-9326/aa9a76

McCarty, J.L., Aalto, J., Paunu, V.V., Arnold, S.R., Eckhardt, S., Klimont, Z., Fain, J.J., Evangeliou, N., Venäläinen, A., Tchebakova, N.M., Parfenova, E.I., Kupiainen, K., Soja, A.J., Huang, L. and **Wilson, S.** (2021) Reviews and syntheses: Arctic fire regimes and emissions in the 21st century. *Biogeosciences*, 18(18): 5053–5083. DOI: https://doi.org/10.5194/bg-18-5053-2021

NASA Goddard Institute for Space Studies (NASA/GISS) (2018) *NASA-GISS GISS-E2.1G model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.7127

NASA Goddard Institute for Space Studies (NASA/GISS) (2019a) *NASA-GISS GISS-E2-2-G model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.7129

NASA Goddard Institute for Space Studies (NASA/GISS) (2019b) *NASA-GISS GISS-E2.1H model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.7128

O'Neill, B.C., Tebaldi, C., van Vuuren, D.P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.F., Lowe, J., Meehl, G.A., Moss, R., Riahi, K. and **Sanderson, B.M.** (2016) The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, 9(9): 3461–3482. DOI: https://doi.org/10.5194/gmd-9-3461-2016

Parding, K.M., Dobler, A., McSweeney, C.F., Landgren, O.A., Benestad, R., Erlandsen, H.B., Mezghani, A., Gregow, H., Räty, O., Viktor, E., El Zohbi, J., Christensen, O.B. and **Loukos, H.** (2020) GCMeval – An interactive tool for evaluation and selection of climate model ensembles. *Climate Services*, 18: 100167. DOI: https://doi.org/10.1016/j.cliser.2020.100167

Reichler, T. and **Kim, J.** (2008) How Well Do Coupled Models Simulate Today's Climate? *Bulletin of the American Meteorological Society*, 89(3): 303–312. DOI: https://doi.org/10.1175/BAMS-89-3-303

Ridley, J., Menary, M., Kuhlbrodt, T., Andrews, M. and **Andrews, T.** (2019a) *MOHC HadGEM3-GC31-LL model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.6109

Ridley, J., Menary, M., Kuhlbrodt, T., Andrews, M. and **Andrews, T.** (2019b) *MOHC HadGEM3-GC31-MM model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.6112

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Loosveldt Tomas, S. and **Zimmermann, K.** (2020) Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview. *Geoscientific Model Development*, 13(3): 1179–1199. DOI: https://doi.org/10.5194/gmd-13-1179-2020

Rong, X. (2019) *CAMS CAMS_CSM1.0 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.9754

Samantaray, A., Mooney, P.A. and **Vivacqua, C.A.** (2023) Bergen Metrics: Composite error metrics for assessing performance of climate models using EURO-CORDEX simulations. *Geoscientific Model Development Discussions*, 2023: 1–31. DOI: https://doi.org/10.5194/gmd-2023-134

Screen, J.A. and **Simmonds, I.** (2010) The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature*, 464: 1334–1337. DOI: https://doi.org/10.1038/nature09051

Seferian, R. (2018) *CNRM-CERFACS CNRM-ESM2-1 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.4068

Seland, Ø., Bentsen, M., Olivié, D.J.L., Toniazzo, T., Gjermundsen, A., Graff, L.S., Debernard, J.B., Gupta, A.K., He, Y., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K.S., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I.H.H., Landgren, O.A., Liakka, J., Moseid, K.O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T. and **Schulz, M.** (2019) *NCC NorESM2-LM model output prepared for CMIP6 CMIP* *historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.8036

Semmler, T., Danilov, S., Rackow, T., Sidorenko, D., Barbi, D., Hegewald, J., Sein, D., Wang, Q. and **Jung, T.** (2018) *AWI AWI-CM1.1MR model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.2686

Sengupta, S. and **Boyle, J.S.** (1998) Using common principal components for comparing GCM simulations. *Journal of Climate*, 11(5): 816–830. DOI: https://doi.org/10.1175/1520-0442(1998)011<0816:UCPCFC>2.0.CO;2

Stevenson, S., Huang, X., Zhao, Y., Di Lorenzo, E., Newman, M., Xu, T. and **Capotondi, A.** (2023) *UCSB E3SM1.0 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.17109

Stouffer, R. (2019) *UA MCM-UA-1-0 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.8888

Swart, N.C., Cole, J.N., Kharin, V.V., Lazare, M., Scinocca, J.F., Gillett, N.P., Anstey, J., Arora, V., Christian, J.R., Jiao, Y., Lee, W.G., Majaess, F., Saenko, O.A., Seiler, C., Seinen, C., Shao, A., Solheim, L., von Salzen, K., Yang, D., Winter, B. and **Sigmond, M.** (2019a) *CCCma CanESM5-CanOE model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.10260

Swart, N.C., Cole, J.N., Kharin, V.V., Lazare, M., Scinocca, J.F., Gillett, N.P., Anstey, J., Arora, V., Christian, J.R., Jiao, Y., Lee, W.G., Majaess, F., Saenko, O.A., Seiler, C., Seinen, C., Shao, A., Solheim, L., von Salzen, K., Yang, D., Winter, B. and **Sigmond, M.** (2019b) *CCCma CanESM5 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.3610

Swart, N.C., Cole, J.N., Kharin, V.V., Lazare, M., Scinocca, J.F., Gillett, N.P., Anstey, J., Arora, V., Christian, J.R., Jiao, Y., Lee, W.G., Majaess, F., Saenko, O.A., Seiler, C., Seinen, C., Shao, A., Solheim, L., von Salzen, K., Yang, D., Winter, B., Sigmond, M., Abraham, C., Akingunola, A. and **Reader, C.** (2019c) *CCCma CanESM5.1 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.17339

Tang, Y., Rumbold, S., Ellis, R., Kelley, D., Mulcahy, J., Sellar, A., Walton, J. and **Jones, C.** (2019) *MOHC UKESM1.0-LL model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.6113

Tatebe, H. and **Watanabe, M.** (2018) *MIROC MIROC6 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.5603

Taylor, K.E. (2001) Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7): 7183–7192. DOI: https://doi.org/10.1029/2000JD900719

Voldoire, A. (2018) *CMIP6 simulations of the CNRM-CERFACS based on CNRM-CM6-1 model for CMIP experiment historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.4066

Voldoire, A. (2019) *CNRM-CERFACS CNRM-CM6-1-HR model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.4067

Volodin, E., Mortikov, E., Gritsun, A., Lykossov, V., Galin, V., Diansky, N., Gusev, A., Kostrykin, S., Iakovlev, N., Shestakova, A. and **Emelina, S.** (2019a) *INM INM-CM4-8 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.5069

Volodin, E., Mortikov, E., Gritsun, A., Lykossov, V., Galin, V., Diansky, N., Gusev, A., Kostrykin, S., Iakovlev, N., Shestakova, A. and **Emelina, S.** (2019b) *INM INM-CM5-0 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.5070

Wieners, K.H., Giorgetta, M., Jungclaus, J., Reick, C., Esch, M., Bittner, M., Legutke, S., Schupfner, M., Wachsmann, F., Gayler, V., Haak, H., de Vrese, P., Raddatz, T., Mauritsen, T., von Storch, J.S., Behrens, J., Brovkin, V., Claussen, M., Crueger, T., Fast, I., Fiedler, S., Hagemann, S., Hohenegger, C., Jahns, T., Kloster, S., Kinne, S., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Müller, W., Nabel, J., Notz, D., Peters-von Gehlen, K., Pincus, R., Pohlmann, H., Pongratz, J., Rast, S., Schmidt, H., Schnur, R., Schulzweida, U., Six, K., Stevens, B., Voigt, A. and **Roeckner, E.** (2019) *MPI-M MPI-ESM1.2-LR model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.6595

Williams, R.S., Marshall, G.J., Levine, X., Graff, L.S., Handorf, D., Johnston, N.M., Karpechko, A.Y., Orr, A., de Berg, W.J.V., Wijngaard, R.R. and **Mooney, P.A.** (2024) Future Antarctic Climate: Storylines of mid-latitude jet strengthening and shift emergent from CMIP6. *Journal of Climate*, 37(7): 2157–2178. DOI: https://doi.org/10.1175/JCLI-D-23-0122.1

Wu, T., Chu, M., Dong, M., Fang, Y., Jie, W., Li, J., Li, W., Liu, Q., Shi, X., Xin, X., Yan, J., Zhang, F., Zhang, J., Zhang, L. and **Zhang, Y.** (2018) *BCC BCC-CSM2-MR model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.2948

Yokoi, S., Takayabu, Y.N., Nishii, K., Nakamura, H., Endo, H., Ichikawa, H., Inoue, T., Kimoto, M., Kosaka, Y., Miyasaka, T., Oshima, K., Sato, N., Tsushima, Y. and **Watanabe, M.** (2011) Application of Cluster Analysis to Climate Model Performance Metrics. *Journal of Applied Meteorology and Climatology*, 50(8): 1666–1675. DOI: https://doi.org/10.1175/2011JAMC2643.1

Yu, Y. (2019) *CAS FGOALS-f3-L model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.3355

Yukimoto, S., Koshiro, T., Kawai, H., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., Yoshimura, H., Shindo, E., Mizuta, R., Ishii, M., Obata, A. and **Adachi, Y.** (2019) *MRI MRI-ESM2.0 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.6842

Zappa, G. and **Shepherd, T.G.** (2017) Storylines of atmospheric circulation change for European regional climate impact assessment. *Journal of Climate*, 30(16): 6561–6577. DOI: https://doi.org/10.1175/JCLI-D-16-0807.1

Ziehn, T., Chamberlain, M., Lenton, A., Law, R., Bodman, R., Dix, M., Wang, Y., Dobrohotoff, P., Srbinovsky, J., Stevens, L., Vohralik, P., Mackallah, C., Sullivan, A., O'Farrell, S. and **Druken, K.** (2019) *CSIRO ACCESS-ESM1.5 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. DOI: https://doi.org/10.22033/ESGF/CMIP6.4272

TO CITE THIS ARTICLE:
Graff, L.S., Landgren, O.A., Parding, K.M., Levine, X., Williams, R.S. and Mooney, P.A. (2026) Selecting CMIP6 Models for Future Arctic Storylines Using a Novel Performance Score. *Tellus A: Dynamic Meteorology and Oceanography* 78(1): 1–21. DOI: https://doi.org/10.16993/tellusa.4111

Submitted: 28 March 2025 **Accepted:** 10 November 2025 **Published:** 02 January 2026

Tellus A: Dynamic Meteorology and Oceanography is a peer-reviewed open access journal published by Stockholm University Press.