Received: 15 April 2024 | Revised: 21 October 2024 | Accepted: 6 January 2025

DOI: 10.1002/asl.1290

Atmospheric Science Letters RMetS

RESEARCH ARTICLE

Spatial and temporal dependence in distribution-based evaluation of CMIP6 daily maximum temperatures

Mala Virdee[1] | **Ieva Kazlauskaite**[2,3] | **Emma J. D. Boland**[4] | **Emily Shuckburgh**[1] | **Alison Ming**[5]

[1]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

[2]Department of Engineering, University of Cambridge, Cambridge, UK

[3]Department of Statistical Science, University College London, London, UK

[4]British Antarctic Survey, Cambridge, UK

[5]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

Correspondence
Mala Virdee, Department of Computer Science and Technology, University of Cambridge, Cambridge, UK.
Email: mv490@cam.ac.uk

Funding information
UK Research and Innovation,
Grant/Award Number: EP/S022961/1

Abstract

Climate models are increasingly used to derive localised, specific information to guide adaptation to climate change. Model projections of future scenarios are conferred credibility by evaluating model skill in reproducing large-scale properties of the observed climate system. Model evaluation at fine spatial and temporal scales and for rare extreme events is critical for provision of reliable adaptation-relevant information, but may be challenging given significant internal variability and limited observed data in this setting. Comparing distributions of physical variables from historical simulations of Coupled Model Intercomparison Project models against observed distributions provides a comprehensive, concise and physically-justified skill measure. Calculating divergence between distributions requires aggregation of data spatially or temporally. The spatial and temporal scales at which a divergence measure converges to a consistent value can indicate the scales at which a well-defined climate signal emerges from internal variability. Below this threshold, there may be insufficient data for robust evaluation, particularly for rare extremes. Here, the behaviour of several divergence measures in response to spatial and temporal aggregation is analysed empirically to give a novel evaluation of CMIP6 daily maximum temperature simulations against reanalysis. Some key insights presented here can inform methodological choices made when deriving adaptation-relevant information. Convergence varies according to model, geographic region and divergence measure; selection of the level of precision at which models can provide reliable information therefore requires a context-specific understanding. For this purpose, an interactive tool provided alongside this study demonstrates scale-dependent evaluation across several geographic regions. Commonly applied measures are found to be only weakly sensitive to discrepancies in the tails of distributions.

KEYWORDS

climate models, CMIP6, model evaluation, temperature extremes

1 | INTRODUCTION

Continuous improvement of General Circulation Models (GCMs) over successive generations of the Coupled Model Intercomparison Project (CMIP) has enabled robust detection and attribution of global climate change and provision of long-term projections of future warming scenarios to guide emissions reduction targets (Bock et al., 2020; Edwards, 2013). Models have been developed, calibrated and validated for their ability to reproduce large-scale average physical properties of the observed climate system. However, GCMs are necessarily imperfect, finite-resolution representational tools of an inherently uncertain target system; the range of questions they may be expected to answer adequately is limited, and these limits are not well established.

Increasing urgency under intensifying climate risk has shifted emphasis towards adaptation in recent decades (Klein et al., 2007; Seneviratne et al., 2021). Impact assessment for adaptation typically requires more precise information at local spatial scales and sector-relevant timescales at which adaptation measures can be implemented and concerning the occurrence of high-impact extremes rather than average quantities (Oreskes et al., 2010). “Climate scientists have an intuitive feeling for [the scales at which models are reliable] and use it when interpreting results” (Masson & Knutti, 2011). However, it is increasingly not only climate scientists interpreting and applying model output but also stakeholders across climate-sensitive sectors concerned with assessing and adapting to risk (Vaughan & Dessai, 2014). This leaves open the possibility that models may be applied to answer questions more specific than those for which robust assessment of their historical skill lends future projections credibility.[1] There is a need to establish best practices when deriving adaptation-relevant predictions from GCMs to ensure the trustworthiness of information provided and to avoid sub-optimal allocation of limited adaptation resources (Nissan et al., 2019). Widely used model evaluation approaches may not be up to date with the needs of users of climate information in the adaptation-driven context. Here, two limitations of existing approaches are highlighted.

First, model evaluation does not routinely consider dependence on spatial and temporal scales. On one hand, precise grid-point, time-step-level comparison of model output against observations gives a misleadingly pessimistic assessment given the dominance of internal variability at this scale, particularly for extremes. GCMs are not weather forecast models that might be expected to provide calibrated, synchronous weather predictions, but rather models that should capture the statistical properties of the climate at some aggregate scale. On the other hand, standard approaches evaluating model output averaged over large scales succeed in extracting a signal from internal variability but do not retain the specificity required by decision-makers. Additionally, the onset and severity of climate impacts are often highly geographically heterogeneous, and effective adaptation information should convey this.

Second, standard evaluation approaches (Seneviratne et al. (2021)) often compare summary statistics (mean, variance or higher-order moments) of model output physical variables to corresponding observed quantities. Perkins et al. (2007), Guttorp (2011) and others assert that summary statistics are insufficient for assessing variability and the underlying processes driving extremes and argue for evaluation based on full simulated distributions. Capturing the full shape of an observed distribution is a more stringent criterion than matching a summary statistic and can therefore warrant greater confidence that a model scores highly for physically justified reasons. A skilled simulator of the mean and variance of a physical variable may not capture other attributes (Kharin & Zwiers, 2000); relationships between changes in the mean and changes in extremes are potentially complex and nonlinear (Mearns et al., 1984). Capturing the observed non-Gaussian tails of surface temperatures (Linz et al., 2018) is important for simulation of future changes in extreme temperatures (Catalano et al., 2020).

To address these two limitations, we propose a scale-dependent evaluation approach using divergence measures to compare the full simulated distributions of physical variables against observational reanalysis. This gives insight into the spatial and temporal scales at which models can be considered reliable. Divergence measures are widely used to quantify discrepancies between distributions. To demonstrate, three divergence measures are used to evaluate CMIP6 daily maximum surface air temperature simulations. Increasing prevalence of temperature extremes in recent years is associated with severe societal hazards including increased human morbidity and mortality (Handmer et al., 2012). The Hellinger, Wasserstein and Integrated Quadratic (IQ) distances are selected to reflect some different possible approaches to distribution comparison (see Appendix B.1). Scale dependence of model skill is studied by aggregating data spatially and temporally.

1.1 | Related work

Few studies evaluating summary statistics of GCM-simulated physical variables have explicitly analysed scale dependence.[2] Papalexiou et al. (2020) evaluate summary statistics of CMIP6 temperatures including higher distributional moments at several temporal scales. Sakaguchi et al. (2012) evaluate spatial and temporal scale-dependence of annual surface air temperature trend in CMIP3 and CMIP5, finding general improvement with scale, abrupt improvement at spatial scales larger than 30° × 30°, and finer-scale improvements between successive CMIP generations. Masson and Knutti (2011) study spatial scale dependence of CMIP3 temperature and precipitation by applying smoothing with a variable-scale parameter. Optimal smoothing scales vary by variable and location, and improvements in model resolution are found to not necessarily yield better agreement with large-scale observed temperatures.

Recent studies increasingly take a distribution-based evaluation approach. Abdelmoaty et al. (2021) assess simulation of CMIP6 regional extreme precipitation distributions using the Hellinger distance, finding that models may capture shape properties of the observed distribution better than mean and variance, and that best-performing models according to assessment of summary statistics may differ from best-performing models according to distribution comparison. Vissio et al. (2020) evaluate CMIP6 temperature, precipitation and sea ice cover using the Wasserstein distance, demonstrating that this approach can help pinpoint model weaknesses. Thorarinsdottir et al. (2020) use the IQ distance to evaluate CMIP5 and CMIP6 temperature extremes, ranking models and providing model selection guidelines. To our knowledge, a distribution-based evaluation that explicitly considers scale dependence has not previously been conducted.

2 | DATA

To demonstrate the proposed methodology, historical simulations of daily maximum surface air temperature from 5 GCMs contributing to the latest model intercomparison phase, CMIP6, are evaluated. These GCMs (GFDL-ESM4, IPSL-CM6A-LR, MPI-ESM1-2-HR, MRI-ESM2-0 and UKESM1-0-LL) are the primary models selected for the Inter-Sectoral Impacts Model Intercomparison Project (ISIMIP) meeting criteria of performance, structural independence and spanning the CMIP6 Equilibrium Climate Sensitivity (ECS) range (Lange, 2021; Warszawski et al., 2014). For each model, daily maximum temperature simulations are taken from one realisation from the historical experiment simulating the recent past period 1850–2014. Model details are included in Table A1.

Reanalyses are constructed by assimilating observations into models to provide gridded, physically consistent data. W5E5 (Lange et al., 2021), treated as the ground truth against which models are evaluated, is a high-resolution reanalysis originally compiled to support bias adjustment of simulations for ISIMIP impact assessments. Hereafter W5E5 is referred to as the reference. A second reanalysis, MERRA2 (Gelaro et al., 2017), is treated as an additional pseudo-model, providing an indication of the best possible model performance that could be expected and giving an estimate of the relative contributions of model error and internal variability. Details of W5E5 and MERRA2 are included in Table A2.

The temporal overlap of these datasets gives a 35-year period spanning 1980 to 2014 for analysis. Bilinear

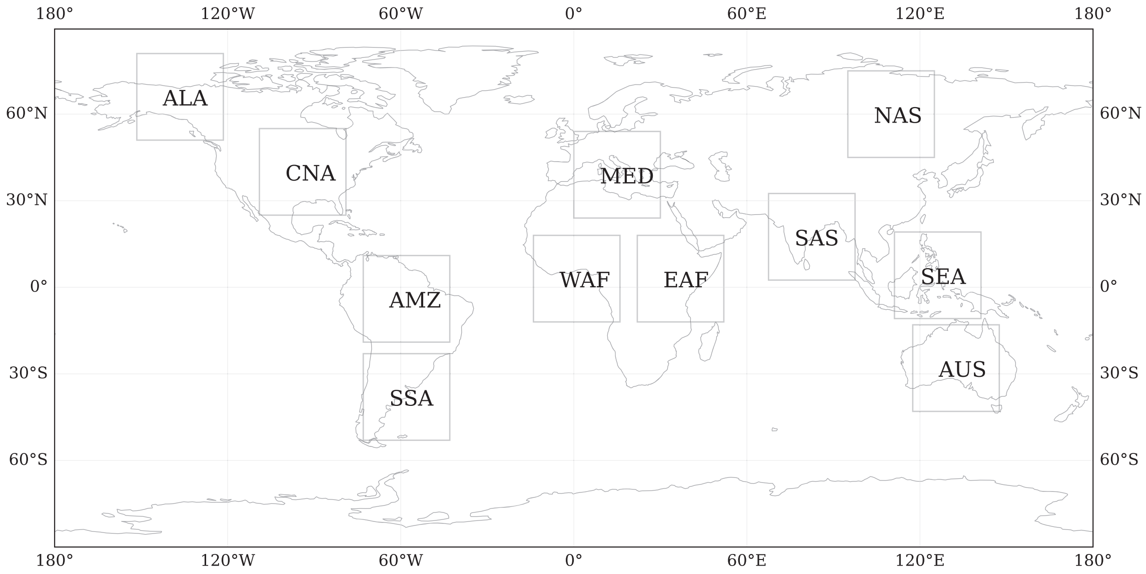


FIGURE 1 Map showing boundaries of 11 30° × 30° latitude × longitude regions chosen for analysis. The full region names and boundaries are included in Table A3.

interpolation of all data onto a 1° × 1° spatial grid is applied to facilitate intercomparison following standard practice in impact-relevant studies (Almazroui et al., 2020; Shi et al., 2018). Eleven 20° × 30° geographic regions, derived from the widely used Giorgi regions (Christensen et al., 2007; Giorgi & Francisco, 2000; Houghton et al., 2001), are selected as shown in Figure 1; region details are included in Table A3. Regional distributions of daily maximum temperature for CMIP6 models, MERRA2 and W5E5 are shown in Figure 2. In

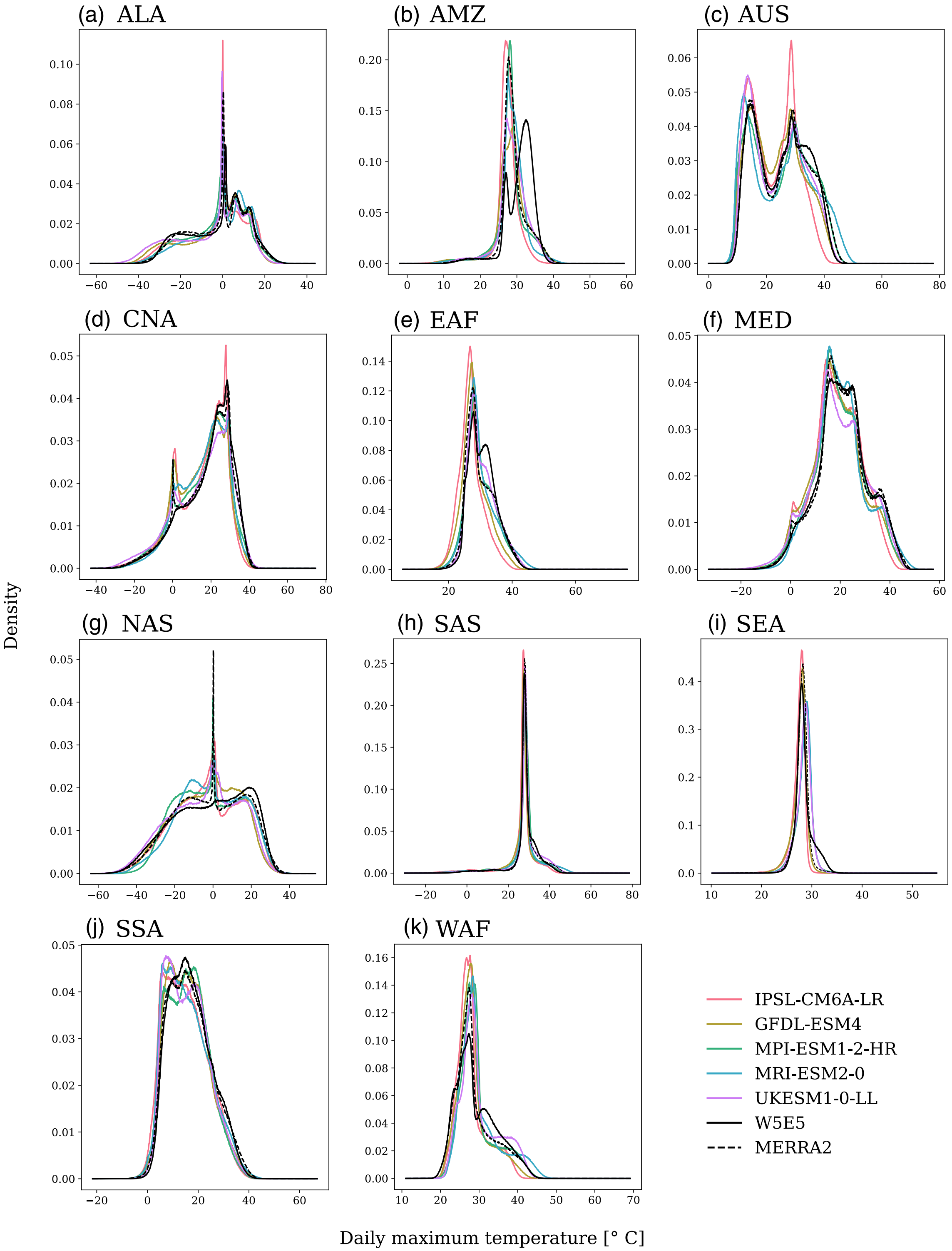


FIGURE 2 Distributions of daily maximum surface temperature simulations from 5 CMIP6 models, MERRA2 reanalysis and W5E5 reanalysis for 11 geographic regions for the period 1980–2014.

some regions, particularly ALA, CNA, MED and NAS, distributions exhibit a 0°C spike. This may reflect a snow-melt-related modelling bias in the land surface model (e.g., see Qiao et al. (2022)); full analysis of this behaviour is beyond the scope of this study.

3 | METHODOLOGY

3.1 | Spatial and temporal aggregation

Figure 2 shows distributions of 1° × 1° daily maximum temperatures aggregated over each 30° × 30° geographic region for the full time period 1980–2014. At these scales, most models broadly capture the shape of the reference distribution. This approach aims to understand how the discrepancy between models and reference evolves as data is aggregated from the most precise grid-point, time-step level up to the full region and time period. Several possible useful notions of spatial and temporal aggregation of data may be defined; in previous studies, the method used may be implicit. As illustrated in Table 1, three methods are intercompared here:

1. *Centred zoom*: Data are aggregated (a) spatially by including $s \times s$ contiguous grid cells from the region centre, or (b) temporally by including t contiguous days from the start date. This enables localisation of model errors and is informative if considering a specific date or location such as a city. However, this method depends on a central point or start date and localised inhomogeneities as data are aggregated – it therefore does not give a generalised indication of error over the region or time period.
2. *Regular subdivision*: Data are divided into increasingly large regular, equally-sized (a) $s \times s$ spatial subsets or (b) t-day time subsets. A model subset is compared against the corresponding reference subset; results are calculated by averaging over subsets. This gives a generalisation of the centred zoom that does not centre on a particular location or date, instead sampling many different localised spatial and temporal contexts and taking into account all data at every aggregation step. However, aggregation steps are irregular as they are limited to regular subdivisions of the data.
3. *Random subdivision*: N increasingly large, equally sized (a) $s \times s$ spatial subsets or (b) t-day time subsets are sampled at random.[3] This gives a generalisation of regular subdivision not limited to regular subdivisions of the data.

3.2 | Metrics

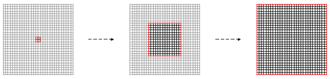
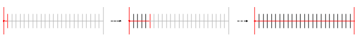
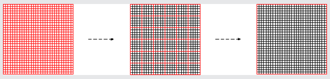
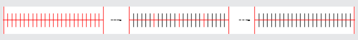
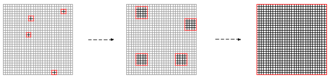
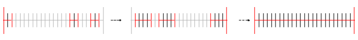
A divergence measure quantifies distance or discrepancy between distributions, a central task within many problems in statistical modelling and machine learning (Markatou et al., 2021). Here, the Hellinger, Wasserstein and Integrated Quadratic (IQ) distances are calculated.[4] Definitions, properties and a review of relevant applications of these measures are included in Appendix B.1. In the absence of a theoretical basis on which to assert the advantages or disadvantages of a particular measure for any given task, it is useful to develop an empirical understanding of their behaviour. For this purpose, the first four L-moments, λ_1 (L-location or mean), λ_2 (L-scale or variance), τ_3 (L-skewness) and τ_4 (L-kurtosis)[5] (Appendix B.2) are also calculated.

4 | RESULTS

4.1 | Spatial aggregation

In this section, the dependence of divergence of daily maximum temperature distributions from the reference on spatial aggregation is considered. Figure 3 shows an

TABLE 1 Three methods of aggregating gridded time series data (a) on a spatial grid and (b) on a timeline into increasingly large subsets.

	(a) Spatial	(b) Temporal
1: Centred zoom		
2: Regular subdivision		
3: Random subdivision		

Note: The black and grey lines indicate (a) spatial grid or (b) time-step of model output. Grey lines indicate data that is not selected. Each red square (a) or red interval (b) represents a data subset which is intercompared with the corresponding subset in the reference data.

example for the Mediterranean Basin (MED). The distribution broadens with spatial aggregation, and distinct peaks in the full regional distribution become better defined (Figure 3a,b). Divergence between MERRA2 and the reference quickly decreases to a small value according to all three measures (Figure 3c–e, dashed black lines). Correspondingly, the *L*-moments for MERRA2 closely match the reference as data are aggregated spatially (Figure 3f–i, dashed black lines). In contrast, for the models (Figure 3c–e, coloured lines), increasing aggregation does not necessarily yield decreasing divergence. Across measures, divergence for GFDL-ESM4 and IPSL-CM6A-LR remains highest with aggregation (yellow and red lines); according to the Wasserstein and IQ distances, divergence for several models begins to increase at $10-15^{\circ}$. Comparison to *L*-moments indicates that GFDL-ESM4 and IPSL-CM6A-LR, which rank lowest across divergence measures, also simulate the mean furthest from the reference (Figure 3f–i, yellow and red lines). The increasing trend in the reference mean at $10-15^{\circ}$ is not well-represented

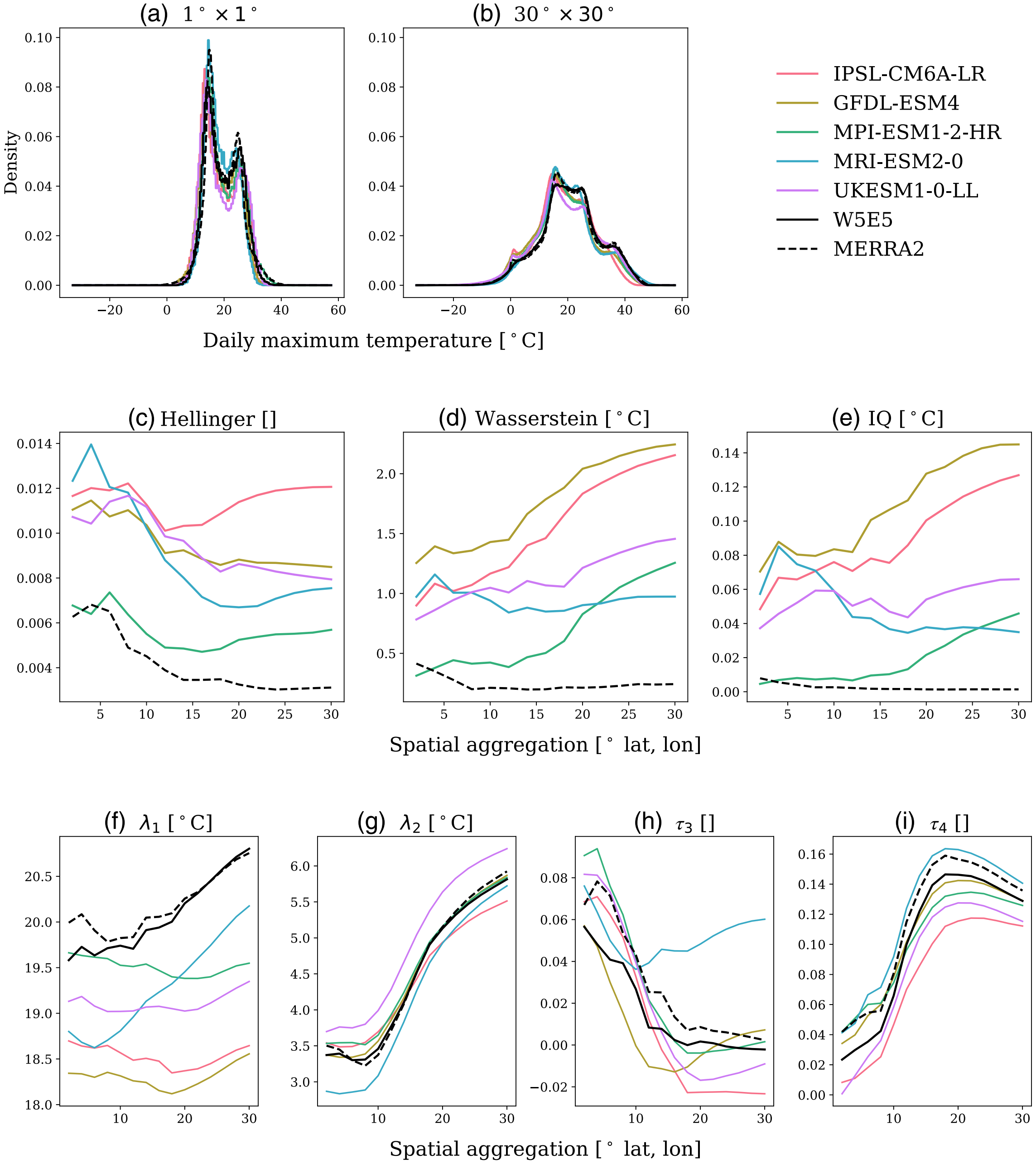


FIGURE 3 MED, spatial centred zoom. (a) and (b): Distributions of daily maximum temperature in the region Mediterranean Basin (MED), aggregated at the smallest (1) and largest (30°) spatial scales respectively using aggregation by centred zoom. (c)–(e): Hellinger, Wasserstein and IQ distance between the model and reference distributions for increasing spatial aggregation by centred zoom. (f)–(i): First four L-moments λ_1 (mean), λ_2 (variance), μ_3 (skewness) and μ_4 (kurtosis) for increasing spatial aggregation.

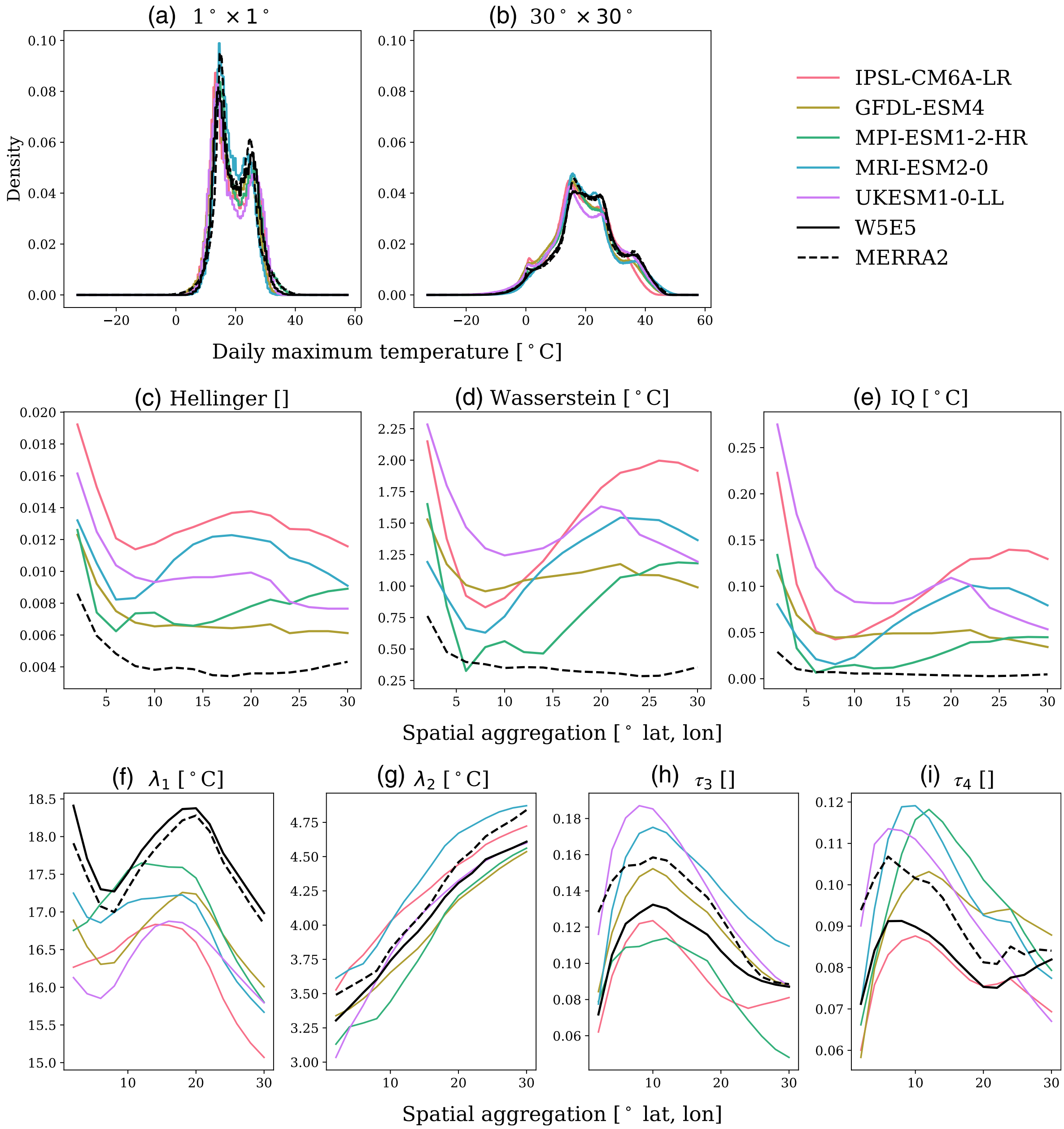


FIGURE 4 SSA, spatial centred zoom. (a) and (b) Distributions of daily maximum temperature in the region Southern South America (SSA), aggregated at the smallest (1°) and largest (30°) spatial scales respectively, using aggregation by centred zoom. (c)–(e): Hellinger, Wasserstein and IQ distance between the model and reference distributions for increasing spatial aggregation by centred zoom. (f)–(i): First four L-moments λ_1 (mean), λ_2 (variance), μ_3 (skewness) and μ_4 (kurtosis) for increasing spatial aggregation.

by models except MRI-ESM2-0 (Figure 3f, blue line). However, MRI-ESM2-0 significantly positively overestimates skewness as aggregation increases (Figure 3h, blue line), but nonetheless performs relatively well according to all three divergence measures (Figure 3c–3e, blue line). This highlights that the divergence measures primarily reward closeness of means whilst insufficiently accounting for discrepancies in higher-order moments.

Figure 4 shows a further example of spatial aggregation by centred zoom for Southern South America (SSA). Figure 4a,b shows the distributions at 1° and 30° aggregation, respectively. As before, divergence between MERRA2 and reference (Figure 4c–e, dashed black line) quickly decreases to a small value, whilst for some models – particularly IPSL-CM6A-LR and MRI-ESM2-0 (Figure 4c–e, red and blue lines) – divergence remains high or begins to increase at $5-10^{\circ}$. Comparison against *L*-moments (Figure 4f–i) again indicates that these measures primarily reflect that these models rank poorly in capturing the fluctuating mean of the reference (Figure 4f), rather than the discrepancy in higher-order moments.

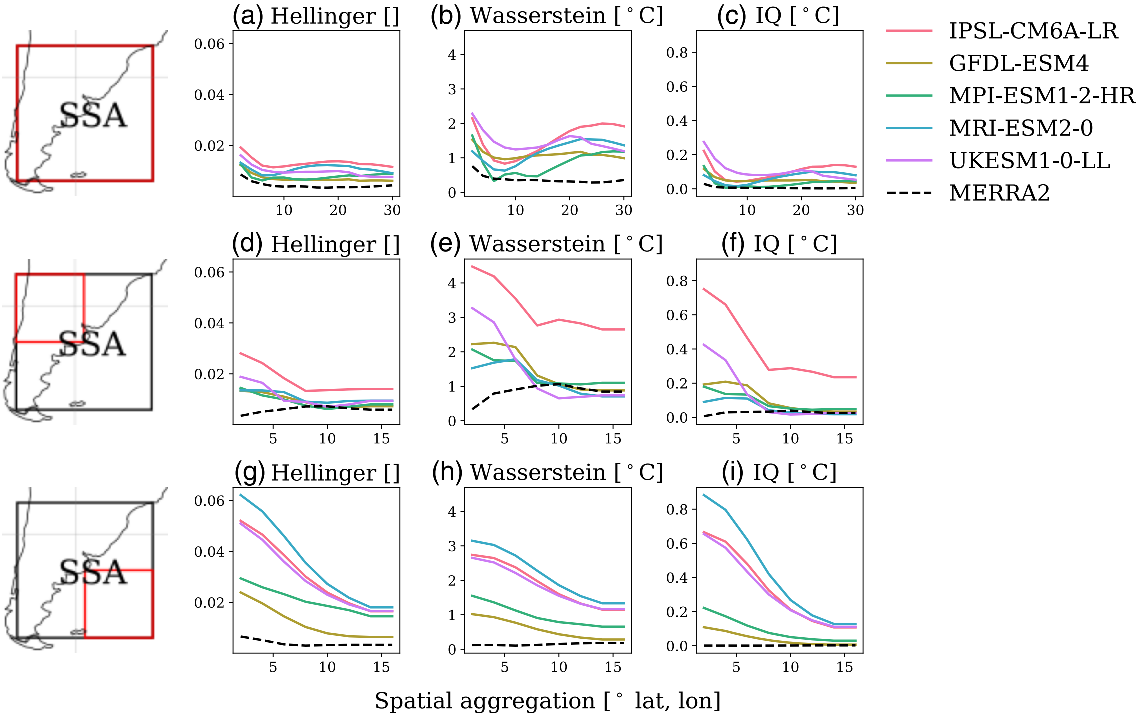


FIGURE 5 SSA sub-regions, spatial centred zoom. Further analysis of the spatial centred zoom aggregation example in Figure 4 for region SSA, subdividing the region into ocean-centred and land-centred sub-regions.

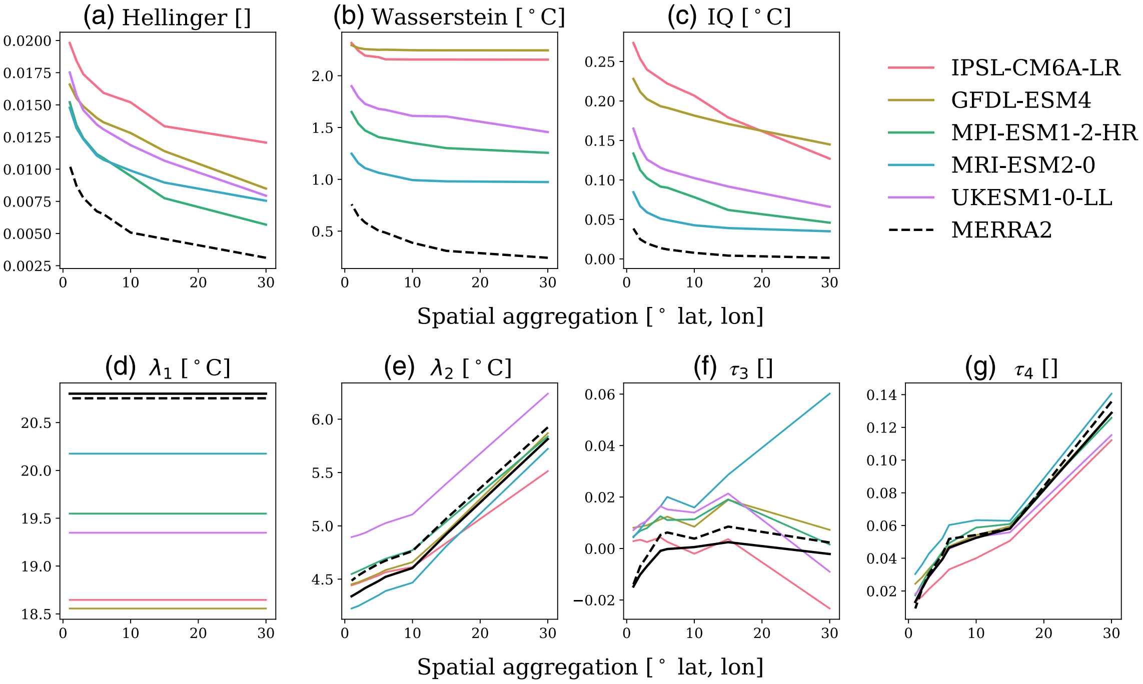


FIGURE 6 MED, spatial regular subdivision. (a)–(c): Hellinger, Wasserstein and IQ distance between the model and reference distributions for increasing spatial aggregation by regular subdivision for region MED. (d)–(g): First four L-moments λ_1 (mean), λ_2 (variance), μ_3 (skewness) and μ_4 (kurtosis) for increasing spatial aggregation.

It is informative to further examine the results for SSA by subdividing the region. In Figure 4, for several models, error initially decreases but begins to increase at 5–10°, particularly IPSL-CM6A-LR (Figure 4c–e, red line). In Figure 5, subdivisions of the SSA region are selected, taking diagonal quarters over ocean or land.

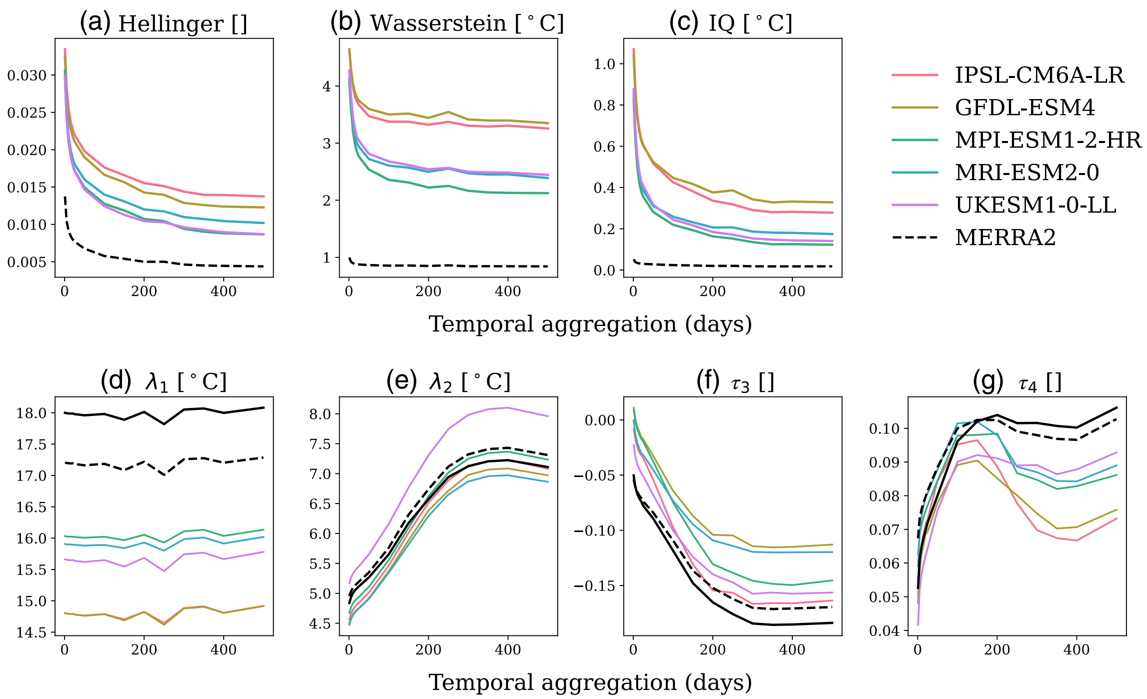


FIGURE 7 CNA, temporal regular subdivision. (a)–(c): Hellinger, Wasserstein and IQ distance between the model and reference distributions for increasing temporal aggregation by regular subdivision for region CNA. (d)–(g): First four L-moments λ_1 (mean), λ_2 (variance), μ_3 (skewness) and μ_4 (kurtosis) for increasing temporal aggregation.

The increase of divergence for some models can be attributed to significant differences in model skill over ocean versus land, most pronounced for IPSL-CM6A-LR (Figure 5, red line). When separated, divergence generally decreases smoothly with aggregation, and decreases more smoothly over ocean. This highlights the utility of the spatially centred zoom approach for diagnosing localised model biases.

Figure 6 shows the analysis for MED, as in Figure 3, but instead using spatial regular subdivision (see Section 3.1). Divergence for models and MERRA2 (Figure 6a–c, coloured lines and black dashed line) decreases with aggregation. As may be expected, the mean remains constant with increasing size of regular subdivisions (Figure 6d), whilst variance and tail heaviness increase (Figure 6e,g). At 30° × 30° aggregation over the full region, regular subdivision and spatially centred zoom (Figure 3) are equivalent. As before, MRI-ESM2-0 positively overestimates skewness as aggregation increases (Figure 6f, blue line) but performs well according to divergence measures (Figures 6a and 3c, blue line). However, localised effects are smoothed as error is averaged over the whole region at every step. Divergence for all models and MERRA2 decreases with aggregation and in many cases converges to a steady value (Figure 6a–c, coloured lines and black dashed line). This indicates the scale at which the data sampled within each subset adequately resolve the reference distributions.

Convergence scale differs across the three measures. The Wasserstein and IQ distances converge by 10–15° for most models, whilst the Hellinger distance continues to decrease. In general, of the three measures, the Wasserstein distance is found to be most sensitive to the mean, which can be estimated with relatively few samples – it tends to converge quickly to a steady value with aggregation. The Hellinger distance is found to be sensitive to outliers and is therefore strongly affected by the reduced contribution of internal variability with aggregation.

Spatial random subdivision (Section 3.1), yields noisier but broadly similar results, omitted here for conciseness, to regular subdivision if the number of samples N is sufficient.

4.2 | Temporal aggregation

In this section, the dependence of divergence of daily maximum temperature distributions from the reference on temporal aggregation is considered. Figures 7 and 8 show examples for Central North America (CNA) and Southeast Asia (SEA) for temporal regular subdivision. Divergence of MERRA2 and all models from the reference (Figure 7a–c and Figure 8a–c) smoothly decreases with aggregation according to all three measures. The temporal scale of convergence varies across measures and between regions. As before, models that capture the reference mean relatively poorly, here GFDL-ESM4 and IPSL-CM6A-LR (Figure 7d and Figure 8d, red and yellow lines), are penalised according to the divergence measures. Meanwhile, discrepancies in higher moments (Figure 7d–g and Figure 8d–g)) are not represented consistently. It is notable in Figures 7e and 8e that variance λ_2 is significantly larger for CNA than SEA (see also 2d and 2i). Additionally, λ_2 increases more rapidly with aggregation in CNA than in SEA, reflecting a stronger seasonal cycle, which is captured as the annual timescale is reached. Each divergence measure converges more quickly for SEA than CNA, as might be expected given a relatively simpler single, strong distributional peak for SEA, whilst greater aggregation may be required to resolve the wider, multi-modal CNA distribution (2d and 2i).

Temporal centred zoom shows fluctuations in divergence with aggregation. However, these effects are sensitive to a start date and less useful for the diagnosis of model errors than the spatial localisation demonstrated in Figure 5. Temporal random subdivision yields noisier but broadly similar results to regular subdivision if the number of samples N is sufficient.

4.3 | Interactive demonstration

Scale-dependent evaluation for all regions and for the three spatial and temporal data aggregation methods proposed in Section 3.1 can be accessed in an interactive tool, as shown in Figure 9 - see Appendix C for details.

5 | DISCUSSION AND CONCLUSIONS

Divergence-based evaluation of climate simulations against reanalysis is sensitive to spatial and temporal scale. This is not routinely considered when deriving adaptation-relevant information. This study yields several methodological recommendations for evaluation and application of GCM simulations of physical variables.

The examples in Sections 4.1 and 4.2, which can be explored further using the interactive tool in Section 4.3, demonstrate dependence of model skill on both aggregation scale and method. Scale-dependent evaluation can enable selection of an appropriate precision level and guide post-processing steps including spatial interpolation and selection of a subset of the CMIP model ensemble, since a ranking of model skill may be scale-dependent.

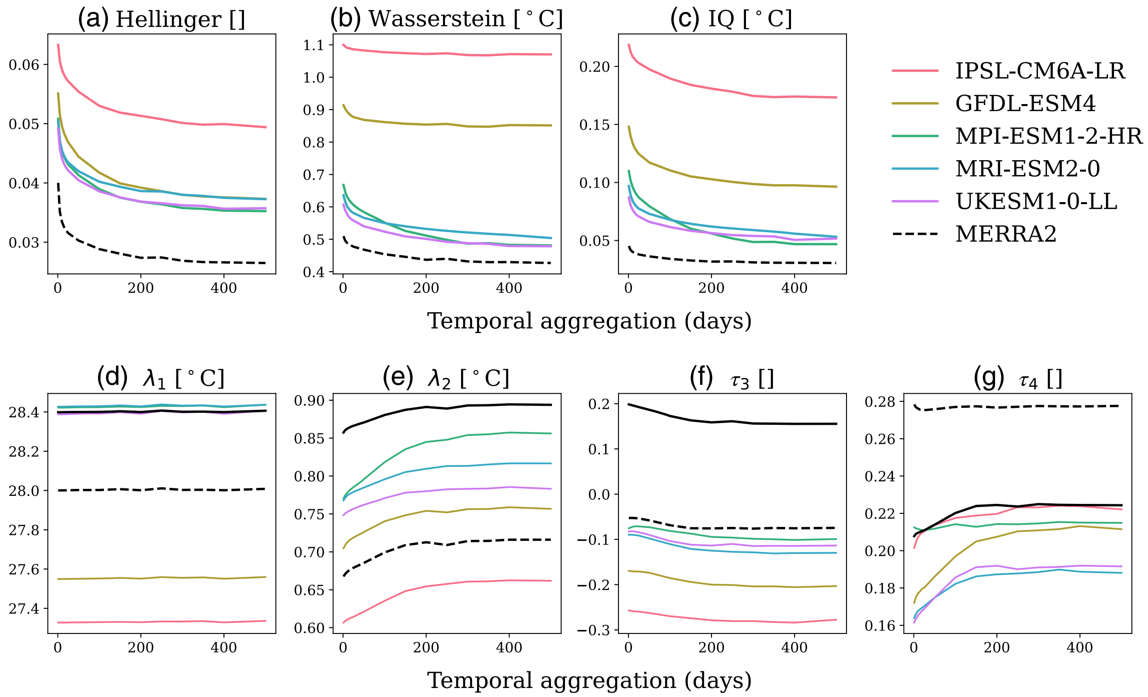
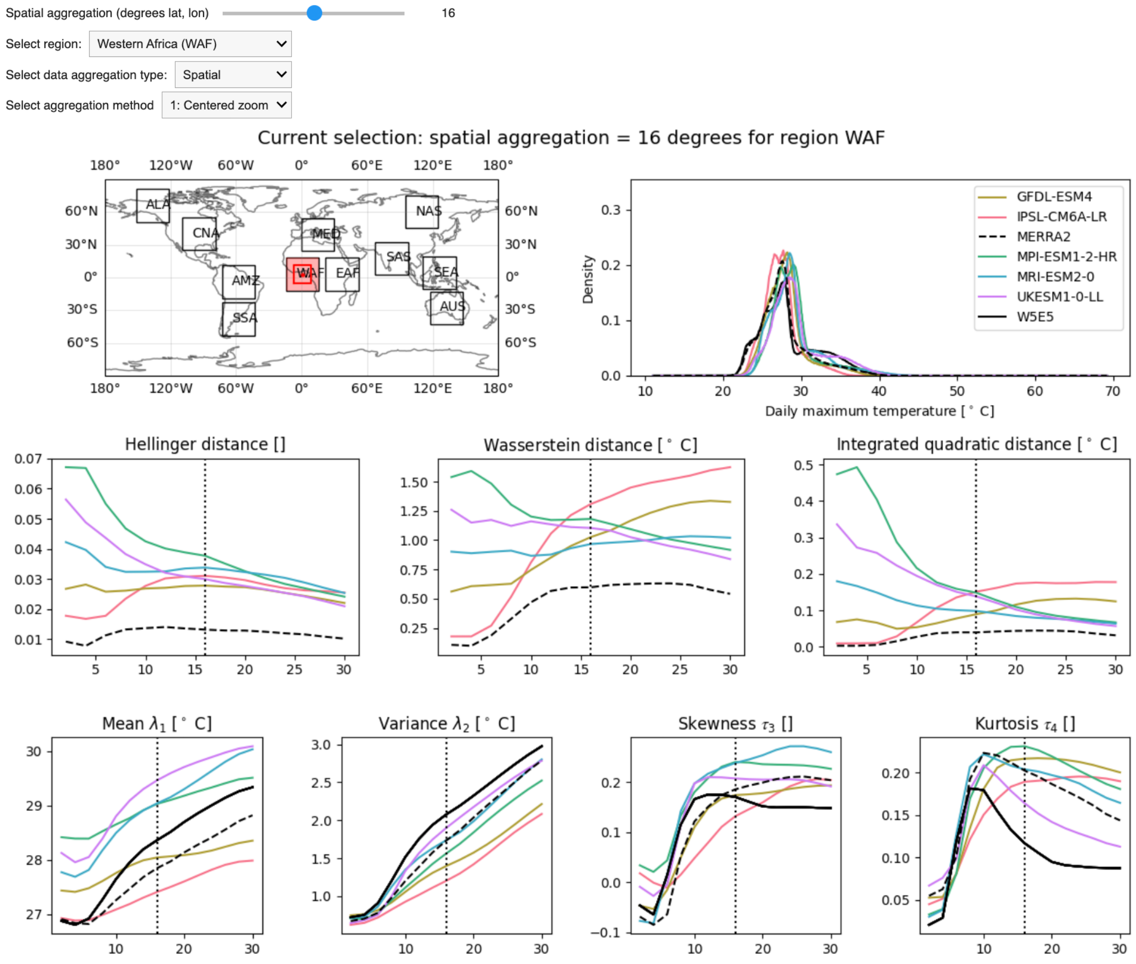


FIGURE 8 SEA, temporal regular subdivision. (a)–(c): Hellinger, Wasserstein and IQ distance between the model and reference distributions for increasing temporal aggregation by regular subdivision for region SEA. (d)–(g): First four L-moments λ_1 (mean), λ_2 (variance), μ_3 (skewness) and μ_4 (kurtosis) for increasing temporal aggregation.



Spatial aggregation (degrees lat, lon) 16

Select region: Western Africa (WAF)

Select data aggregation type: Spatial

Select aggregation method 1: Centered zoom

Current selection: spatial aggregation = 16 degrees for region WAF

FIGURE 9 Screenshot of interactive demonstration for all regions and data aggregation methods, available to access online.

The three aggregation methods defined in Section 3.1 answer different questions; the most appropriate method depends on the nature of the task. Centred zoom aids diagnosis of spatially localised model errors, particularly where an increase in error with aggregation is observed. This may be apt if a studying central point such as a city – the scale of aggregation required for model skill to improve adequately and can be determined in a way that is tailored to the location in question. It may also aid the definition of geographic boundaries for regional analysis by highlighting where results may be impacted by localised errors or the inclusion of different neighbouring climate regimes.

Regular subdivision can indicate scales at which representative distributions emerge from internal variability within a particular geographic region. This gives a means of answering a critical question not often asked when post-processing GCM output: at what scales are models reliable? Divergence is typically high for small spatial and temporal scales where internal variability dominates and converges to a steady value with aggregation. Convergence scale varies by model and region and is sensitive to choice of divergence measure. It may be misleading to use climate models to derive information below the scales at which they can be robustly evaluated and below which the evaluative measure would be dominated by localised effects. Regional intercomparison indicates convergence typically occurs more rapidly for regions where the distribution of daily maximum temperatures is more normal – in other words, in regions where temperature is more homogeneous. Therefore, if conducting a regional analysis, it may be important to consider that strong seasonal cycles and inhomogeneity of geographic features give rise to more complex distributions that may require a larger spatial or temporal scale to be adequately resolved.

Intercomparison of three divergence measures tested gives several insights. Firstly, divergence-based evaluation is sensitive to choice of measure; in several examples (Figure 6) skill-based model ranking would change based on divergence measure as well as scale. It may be informative to test sensitivity to choice of measure during model evaluation. Secondly, the measures are only weakly sensitive to discrepancies in higher-order moments relevant for assessing tails of distributions. Commonly used measures may be inadequate if the priority is assessing extremes. Development and application of measures such as kernel-based divergences which allow emphasis to be placed on particular parts of distributions is a potential direction for future research.

A limitation of this study is the use of reanalysis as ground truth for model evaluation. Reanalyses contain compound errors from imperfections of data assimilation, model uncertainty, and limited availability of input data outside well-observed regions and beyond the recent past. Errors are likely shared across reanalyses and particularly affect the representation of extremes (Bosilovich et al., 2013; Donat et al., 2014). Daily maximum temperature is relatively well-represented in reanalyses; these limitations may therefore present a greater challenge when evaluating other physical variables such as precipitation.

Areas for future work include evaluation of other physical variables – including those where distributions are less normal and behaviour with aggregation is more complex – and multivariate analysis. Aggregation of multiple realisations from each GCM could be considered, particularly given the growing availability of large single-model initial conditions ensembles designed to estimate internal variability. Whilst this study did not include a comprehensive evaluation or ranking of the CMIP6 ensemble, future work could determine whether particular models consistently perform best within a scale-dependent evaluation framework. Analysis enabling generalised conclusions about the sensitivity of divergence measures to distributional shape and complexity is another valuable direction for future consideration.

AUTHOR CONTRIBUTIONS

Mala Virdee: Conceptualization; writing – original draft; investigation; methodology; writing – review and editing; formal analysis; visualization. **Ieva Kazlauskaite:** Conceptualization; investigation; writing – review and editing; methodology; formal analysis; supervision. **Emma J. D. Boland:** Conceptualization; investigation; methodology; writing – review and editing; formal analysis; supervision. **Emily Shuckburgh:** Conceptualization; investigation; writing – review and editing; methodology; formal analysis; supervision. **Alison Ming:** Supervision; formal analysis; writing – review and editing; methodology; conceptualization; investigation.

FUNDING INFORMATION

This work was supported by the UKRI Centre for Doctoral Training in Application of Artificial Intelligence to the study of Environmental Risks (reference EP/S022961/1).

DATA AVAILABILITY STATEMENT

The data used in this study are openly available: CMIP6 data can be accessed in the Earth System Grid Federation (ESGF) Data Portal via the following URL: https://esgf-index1.ceda.ac.uk/projects/cmip6-ceda/. W5E5 data are available at http://doi.org/10.48364/ISIMIP.342217. MERRA2 data are available at http://doi.org/10.5067/9SC1VNTWGWV3. Code and instructions to reproduce the results presented in this paper can be accessed at https://github.com/mvirdee/divergence_metrics. An interactive tool can be accessed in a Google Collaboratory notebook at https://colab.research.google.com/drive/1DazaztOwLCdaIvonP944UEULc11yha-T.

ORCID

Mala Virdee https://orcid.org/0009-0004-3896-3272
Ieva Kazlauskaite https://orcid.org/0000-0001-9690-0887

ENDNOTES

[1] Further complications are added by a growing range of post-processing methods (e.g., bias correction, statistical and dynamical downscaling – added precision may be misleading without careful evaluation of underlying models (Maraun et al., 2017; Stainforth et al., 2007).

[2] For a review of model evaluation methods used in the context of climate impact modelling, see Raju and Kumar (2020).

[3] For illustration purposes, Table 1 shows $N=4$; in practice, (a) $N=\frac{S^2}{s^2}$ or $N=\frac{T}{t}$ samples are used at each step to give for equivalence to Method 2, where $S \times S$ is the full region size and T is the full time period length. Note that overlapping samples may be selected.

[4] Whilst the Wasserstein and IQ distances can both be interpreted in physical units of degrees Celsius, they encode different concepts of divergence; intercomparison of relatively larger or smaller values for different scales or regions is informative rather than direct comparison of values. The Hellinger distance is dimensionless, taking values between 0 and 1.

[5] L-moment ratios, $\tau_r=\frac{\lambda_r}{\lambda_2}, r=3,4$, give dimensionless measures of skewness and kurtosis independent of scale and satisfy $\tau_r<1$.

[6] A metric has the properties of non-negativity: $d(\mathbb{P},\mathbb{Q})\geq 0$; symmetry: $d(\mathbb{P},\mathbb{Q})=d(\mathbb{Q},\mathbb{P})$; obeying the triangle inequality $d(\mathbb{P},\mathbb{M})\leq d(\mathbb{P},\mathbb{Q})+d(\mathbb{Q},\mathbb{M})$ and obeying the identity of indiscernibles: $d(\mathbb{P},\mathbb{Q})=0$ if and only if $\mathbb{P}=\mathbb{Q}$

[7] A scoring rule s for a probabilistic forecast $\mathbb{A}$ of an event x is said to be proper if $s(\mathbb{A},\mathbb{A})\leq s(\mathbb{A},\mathbb{B})$ for all possible forecasts $\mathbb{A},\mathbb{B}$; the scoring rule is strictly proper if the equality occurs if and only if $\mathbb{A}=\mathbb{B}$. Note that here s is defined such that a smaller score indicates a better forecast. In other words, if a forecaster predicts distribution $\mathbb{A}$, the scoring gives no incentive to quote an alternative distribution $\mathbb{B}$.

[8] [?] refer to the IQ distance as the Cramér distance.

REFERENCES

Abdelmoaty, H.M., Papalexiou, S.M., Rajulapati, C.R. & AghaKouchak, A. (2021) Biases beyond the mean in CMIP6 extreme precipitation: a global investigation. *Earth's Future*, 9(10), e2021EF002196.

Almazroui, M., Saeed, F., Saeed, S., Nazrul Islam, M., Ismail, M., Klutse, N.A.B. et al. (2020) Projected change in temperature and precipitation over Africa from CMIP6. *Earth Systems and Environment*, 4, 455–475.

Arnold, H., Moroz, I. & Palmer, T. (2013) Stochastic parametrizations and model uncertainty in the Lorenz'96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1991), 20110479.

Bellemare, M.G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S. et al. (2017) The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv: 1705.10743*.

Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C. et al. (2020) Quantifying progress across different CMIP phases with the ESMValTool. *Journal of Geophysical Research: Atmospheres*, 125(21), e2019JD032321.

Bosilovich, M.G., Kennedy, J., Dee, D., Allan, R. & O'Neill, A. (2013) On the reprocessing and reanalysis of observations for climate. *Climate Science for Serving Society: Research, Modeling and Prediction Priorities*, 51–71.

Boucher, O., Servonnat, J., Albright, A.L., Aumont, O., Balkanski, Y., Bastrikov, V. et al. (2020) Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, 12(7), e2019MS002010.

Catalano, A.J., Loikith, P.C. & Neelin, J.D. (2020) Evaluating CMIP6 model fidelity at simulating non-Gaussian temperature distribution tails. *Environmental Research Letters*, 15(7), 074026.

Christensen, J., Hewitson, B., Busuioc, A., Chen, A., Gao, X., Held, I. et al. (2007) Regional climate projections. In: *Climate change 2007: the physical science basis*. Cambridge, UK and New York, NY: Cambridge University Press, pp. 847–940.

Csiszár, I. (1967) On information-type measure of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2, 299–318.

Donat, M.G., Sillmann, J., Wild, S., Alexander, L.V., Lippmann, T. & Zwiers, F.W. (2014) Consistency of temperature and precipitation extremes across various global gridded in situ and reanalysis datasets. *Journal of Climate*, 27(13), 5019–5035.

Dunne, J.P., Horowitz, L., Adcroft, A., Ginoux, P., Held, I., John, J. et al. (2020) The GFDL earth system model version 4.1 (GFDL-ESM 4.1): overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, 12(11), e2019MS002015.

Edwards, P.N. (2013) *A vast machine: computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.

Ferson, S., Oberkampf, W.L. & Ginzburg, L. (2008) Model validation and predictive capability for the thermal challenge problem. *Computer Methods in Applied Mechanics and Engineering*, 197(29–32), 2408–2430.

Gardner, P., Lord, C. & Barthorpe, R.J. (2018) An evaluation of validation metrics for probabilistic model outputs. In: *Verification and validation*, Vol. 40795. Minneapolis, MN: American Society of Mechanical Engineers, p. V001T06A001.

Gelaro, R., McCarty, W., Suárez, M.J., Todling, R., Molod, A., Takacs, L. et al. (2017) The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, 30(14), 5419–5454.

Ghil, M. (2016) A mathematical theory of climate sensitivity or, how to deal with both anthropogenic forcing and natural variability? In: *Climate change: Multidecadal and beyond*. Singapore: World Scientific, pp. 31–51. Available from: https://doi.org/10.1142/9789814579933_0002

Giorgi, F. & Francisco, R. (2000) Evaluating uncertainties in the prediction of regional climate change. *Geophysical Research Letters*, 27(9), 1295–1298.

Gneiting, T. & Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Guttorp, P. (2011) The role of statisticians in international science policy. *Environmetrics*, 22(7), 817–825.

Handmer, J., Honda, Y., Kundzewicz, Z.W., Arnell, N., Benito, G., Hatfield, J. et al. (2012) Changes in impacts of climate extremes: human systems and ecosystems. In: *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, NY: Cambridge University Press (CUP), pp. 231–290.

Hosking, J.R. (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 52(1), 105–124.

Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Dai, X. et al. (2001) *Climate change 2001: the scientific basis*, Vol. 881. Cambridge: Cambridge University Press.

Kharin, V.V. & Zwiers, F.W. (2000) Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere–ocean GCM. *Journal of Climate*, 13(21), 3760–3788.

Klein, R.J., Huq, S., Denton, F., Downing, T.E., Richels, R.G., Robinson, J.B. et al. (2007) Inter-relationships between adaptation and mitigation.

Lange, S. (2021) ISIMIP3 bias adjustment fact sheet. Technical Report.

Lange, S., Menz, C., Gleixner, S., Cucchi, M., Weedon, G.P., Amici, A. et al. (2021) WFDE5 over land merged with era5 over the ocean (W5E5 v2. 0).

Linz, M., Chen, G. & Hu, Z. (2018) Large-scale atmospheric control on non-gaussian tails of midlatitude temperature distributions. *Geophysical Research Letters*, 45(17), 9141–9149.

Maraun, D., Shepherd, T.G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J.M. et al. (2017) Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, 7(11), 764–773.

Markatou, M., Karlis, D. & Ding, Y. (2021) Distance-based statistical inference. *Annual Review of Statistics and Its Application*, 8, 301–327.

Masson, D. & Knutti, R. (2011) Spatial-scale dependence of climate model performance in the CMIP3 ensemble. *Journal of Climate*, 24(11), 2680–2692.

Maupin, K.A., Swiler, L.P. & Porter, N.W. (2018) Validation metrics for deterministic and probabilistic data. *Journal of Verification, Validation, and Uncertainty Quantification*, 3(3), 031002.

Mearns, L.O., Katz, R.W. & Schneider, S.H. (1984) Extreme high-temperature events: changes in their probabilities with changes in mean temperature. *Journal of Applied Meteorology and Climatology*, 23(12), 1601–1613.

Müller, A. (1997) Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2), 429–443.

Müller, W.A., Jungclaus, J.H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R. et al. (2018) A higher-resolution version of the max planck institute earth system model (MPI-ESM1. 2-HR). *Journal of Advances in Modeling Earth Systems*, 10(7), 1383–1413.

Nissan, H., Goddard, L., de Perez, E.C., Furlow, J., Baethgen, W., Thomson, M.C. et al. (2019) On the use and misuse of climate change projections in international development. *Wiley Interdisciplinary Reviews: Climate Change*, 10(3), e579.

Oreskes, N., Stainforth, D.A. & Smith, L.A. (2010) Adaptation to global warming: do climate models tell us what we need to know? *Philosophy of Science*, 77(5), 1012–1028.

Papalexiou, S.M., Rajulapati, C.R., Andreadis, K.M., Foufoula-Georgiou, E., Clark, M.P. & Trenberth, K.E. (2021) Probabilistic evaluation of drought in CMIP6 simulations. *Earth's Future*, 9(10), e2021EF002150.

Papalexiou, S.M., Rajulapati, C.R., Clark, M.P. & Lehner, F. (2020) Robustness of CMIP6 historical global mean temperature simulations: trends, long-term persistence, autocorrelation, and distributional shape. *Earth's Future*, 8(10), e2020EF001667.

Perkins, S., Pitman, A., Holbrook, N.J. & McAneney, J. (2007) Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *Journal of Climate*, 20(17), 4356–4376.

Qiao, L., Zuo, Z. & Xiao, D. (2022) Evaluation of soil moisture in cmip6 simulations. *Journal of Climate*, 35(2), 779–800.

Raju, K.S. & Kumar, D.N. (2020) Review of approaches for selection and ensembling of gcms. *Journal of Water and Climate Change*, 11(3), 577–599.

Royston, P. (1992) Which measures of skewness and kurtosis are best? *Statistics in Medicine*, 11(3), 333–343.

Sakaguchi, K., Zeng, X. & Brunke, M.A. (2012) The hindcast skill of the CMIP ensembles for the surface air temperature trend. *Journal of Geophysical Research: Atmospheres*, 117(D16). Available from: https://doi.org/10.1029/2012JD017765

Sánchez de Cos, C., Sánchez-Laulhé, J.M., Jiménez-Alonso, C., Sancho-Avila, J.M. & Rodriguez-Camino, E. (2013) Physically based evaluation of climate models over the Iberian Peninsula. *Climate Dynamics*, 40, 1969–1984.

Sankarasubramanian, A. & Srinivasan, K. (1999) Investigation and comparison of sampling properties of L-moments and conventional moments. *Journal of Hydrology*, 218(1–2), 13–34.

Sellar, A.A., Jones, C.G., Mulcahy, J.P., Tang, Y., Yool, A., Wiltshire, A. et al. (2019) UKESM1: description and evaluation of the UK Earth System Model. *Journal of Advances in Modeling Earth Systems*, 11(12), 4513–4558.

Seneviratne, S.I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A. et al. (2021) Weather and climate extreme events in a changing climate (Chapter 11).

Shi, C., Jiang, Z.-H., Chen, W.-L. & Li, L. (2018) Changes in temperature extremes over China under 1.5 C and 2 C global warming targets. *Advances in Climate Change Research*, 9(2), 120–129.

Silva Lomba, J. & Fraga Alves, M.I. (2020) L-moments for automatic threshold selection in extreme value analysis. *Stochastic Environmental Research and Risk Assessment*, 34(3–4), 465–491.

Simolo, C., Brunetti, M., Maugeri, M., Nanni, T. & Speranza, A. (2010) Understanding climate change–induced variations in daily temperature distributions over Italy. *Journal of Geophysical Research: Atmospheres*, 115(D22). Available from: https://doi.org/10.1029/2010JD014088

Stainforth, D.A., Allen, M.R., Tredger, E.R. & Smith, L.A. (2007) Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857), 2145–2161.

Thorarinsdottir, T.L., Gneiting, T. & Gissibl, N. (2013) Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1), 522–534.

Thorarinsdottir, T.L., Sillmann, J., Haugen, M., Gissibl, N. & Sandstad, M. (2020) Evaluation of CMIP5 and CMIP6 simulations of historical surface air temperature extremes using proper evaluation methods. *Environmental Research Letters*, 15(12), 124041.

Vaughan, C. & Dessai, S. (2014) Climate services for society: origins, institutional arrangements, and design elements for an evaluation framework. *Wiley Interdisciplinary Reviews: Climate Change*, 5(5), 587–603.

Villani, C. et al. (2009) *Optimal transport: old and new*, Vol. 338. Berlin, Heidelberg: Springer.

Vissio, G., Lembo, V., Lucarini, V. & Ghil, M. (2020) Evaluating the performance of climate models based on Wasserstein distance. *Geophysical Research Letters*, 47(21), e2020GL089385.

Vrac, M. & Friederichs, P. (2015) Multivariate—intervariable, spatial, and temporal—bias correction. *Journal of Climate*, 28(1), 218–237.

Wallis, J.R., Matalas, N.C. & Slack, J.R. (1974) Just a moment! *Water Resources Research*, 10(2), 211–219.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O. & Schewe, J. (2014) The inter-sectoral impact model intercomparison project (ISI–MIP): project framework. *Proceedings of the National Academy of Sciences*, 111(9), 3228–3232.

Yuan, Q., Thorarinsdottir, T.L., Beldring, S., Wong, W.K., Huang, S. & Xu, C.-Y. (2019) New approach for bias correction and stochastic downscaling of future projections for daily mean temperatures to a high-resolution grid. *Journal of Applied Meteorology and Climatology*, 58(12), 2617–2632.

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S. et al. (2019) The Meteorological Research Institute Earth System Model version 2.0, MRI-ESM2. 0: description and basic evaluation of the physical component. *Journal of the Meteorological Society of Japan*, 97(5), 931–965.

How to cite this article: Virdee, M., Kazlauskaite, I., Boland, E. J. D., Shuckburgh, E., & Ming, A. (2025). Spatial and temporal dependence in distribution-based evaluation of CMIP6 daily maximum temperatures. *Atmospheric Science Letters*, *26*(2), e1290. https://doi.org/10.1002/asl.1290

APPENDIX A: DATA

A.1 | CMIP6 models

TABLE A1 CMIP6 models and properties.

Model	Reference	Modelling institute, country	Atmosphere resolution (°)	Ocean resolution (°)	Realisation
GFDL-ESM4	Dunne et al. (2020)	Geophysical Fluid Dynamics Laboratory, USA	1.0°	0.4°	r1i1p1f1
IPSL-CM6A-LR	Boucher et al. (2020)	Institut Pierre-Simon Laplace, France	1.8°	0.7°	r1i1p1f1
MPI-ESM1-2-HR	Müller et al. (2018)	Max Planck Institute for Meteorology, Germany	0.9°	0.4°	r1i1p1f1
MRI-ESM2-0	Yukimoto et al. (2019)	Meteorological Research Institute, Japan	1.1°	0.7°	r1i1p1f1
UKESM1-0-LL	Sellar et al. (2019)	Met Office Hadley Centre, UK	1.5°	0.7°	r1i1p1f2

Note: Resolutions are calculated from number of latitude N_{lat} and number of longitude N_{lon} points in grid as $\left(\frac{360}{N_{lon}}\frac{180}{N_{lat}}\right)^{0.5}$. Note that resolutions are nominal and vary with latitude.

A.2 | Reanalysis

W5E5 combines WFDE5 (WATCH (Water and Global Change project) Forcing Data methodology applied to European Centre for Medium Range Weather Forecasting Reanalysis, version 5 (ERA5) reanalysis data over land, with data from ERA5 over the ocean. MERRA2 is the Modern-Era Retrospective analysis for Research and Applications, version 2. Details and references for W5E5 and MERRA2 are shown in Table A2.

TABLE A2 Reanalysis datasets and properties.

Reanalysis	Reference	Spatial resolution	Temporal coverage
W5E5	Lange et al. (2021)	0.5° × 0.5°	1979–2019
MERRA2	Gelaro et al. (2017)	0.5° × 0.625°	1980–present

A.3 | Regions

Abbreviation	Region	West	East	North	South
ALA	Alaska	−151	−121	81	51
AMZ	Amazon Basin	−73	−43	11	−19
AUS	Australia	117	147	−13	−43
CNA	Central North America	−109	−79	55	25
EAF	Eastern Africa	22	52	18	−12
MED	Mediterranean Basin	0	30	54	24
NAS	North Asia	95	125	75	45
SAS	South Asia	67	97	32	2
SEA	Southeast Asia	141	111	19	−11
SSA	Southern South America	−73	−43	−23	−53
WAF	Western Africa	−14	16	18	−12

TABLE A3 Names, abbreviations and boundaries of 11 geographic regions used for analysis.

APPENDIX B: METRICS AND MOMENTS

B.1 | Metrics

A divergence measure $d(\mathbb{P},\mathbb{Q})$ is used to quantify the distance or discrepancy between probability measures $\mathbb{P}$ and $\mathbb{Q}$. This is a central task within many problems in statistical modelling and machine learning (Markatou et al., 2021). Many notions of distance between probability measures have been developed and applied in various contexts; the choice of metric should be based on consideration of the requirements of a task. Gardner et al. (2018), Maupin et al. (2018), Ferson et al. (2008) and others suggest some desirable properties for a divergence measure to be used to validate probabilistic predictions against data, which are useful to consider before introducing the measures used in this analysis.

A divergence measure used for validation should be a metric in the formal sense.[6] It should in some sense generalise the comparison between scalar values, such that the divergence between point probability measures should reduce to the absolute distance $d(\delta_{\mathbb{P}},\delta_{\mathbb{Q}}) = |\delta_{\mathbb{P}} - \delta_{\mathbb{Q}}|$. It should reflect differences in the full statistical distributions of predictions and data, not just the mean and variance, and should have a task-appropriate level of sensitivity to tails without sacrificing robustness. A metric which is expressed in physical units – for instance, a metric having units of degrees Celsius if the prediction and data are in degrees Celsius – can be more intuitive to interpret. Along similar lines, unbounded metrics are able to grow to an arbitrarily large value as discrepancy increases, whereas some metrics are bounded or normalised such that $0 \le d(\mathbb{P},\mathbb{Q}) \le 1$; boundedness or unboundedness may variously be useful for interpretability and intercomparison in different contexts. Assessing the quality of a probabilistic forecast of an event, for instance in probabilistic weather forecasting, is often concerned with the notion of propriety of scoring rules[7] (Gneiting & Raftery, 2007). Thorarinsdottir et al. (2013) extend the notion of proper scoring rules to divergence metrics between probability measures and argue that assessment of climate model skill should use a score divergence which obeys these rules. In the following sections, the properties of three divergence metrics applied in this analysis are briefly outlined in light of these considerations.

B.1.1. | Hellinger distance

The *f*-divergences (Csiszár, 1967) are a family of divergences based on the ratio between probability measures $\mathbb{P}$ and $\mathbb{Q}$ with the general form:

$$d_f(\mathbb{P},\mathbb{Q}) = \int \phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}.$$

Different *f*-divergences are obtained depending on the choice of the function ϕ, notably including the Kullback–Leibler (KL) divergence, Hellinger distance and Total Variation Distance. The Hellinger distance d_{H} between $\mathbb{P}$ and $\mathbb{Q}$ is

$$d_H(\mathbb{P},\mathbb{Q}) = \left(\frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx\right)^{\frac{1}{2}},$$

or, for discrete probability measures:

$$d_H(\mathbb{P},\mathbb{Q})=\left(\frac{1}{2}\sum_{i=1}^{c}\left(\sqrt{p_i}-\sqrt{q_i}\right)^2\right)^{\frac{1}{2}},$$

with a value of zero indicating exact agreement between $\mathbb{P}$ and $\mathbb{Q}$. The Hellinger distance is a metric in the formal sense and has the property of boundedness $0\leq d_H(\mathbb{P},\mathbb{Q})\leq 1$. However, Thorarinsdottir et al. (2013) note that it fails to be a score divergence obeying proper scoring rules.

In previous studies, Arnold et al. (Arnold et al., 2013) use the Hellinger distance for verification of the probabilistic climate predictions from a stochastically parameterised forecast model, finding that it gives a smoother measure of climatological skill than a summary statistic. The Hellinger distance is selected for its robustness and interpretability by Papalexiou et al. (2021) and Abdelmoaty et al. (2021) to evaluate historical CMIP6 extreme precipitation and drought simulations, respectively. It has also been used to develop a physically based evaluation approach assessing the relation between land surface variables simulated by a regional climate model against the corresponding relation in reanalysis (Sánchez de Cos et al., 2013).

B.1.2. | Wasserstein distance

Integral Probability Metrics (IPMs) (Müller, 1997) are a class of divergence measures based on finding the maximum difference between expectations over a space of continuous functions $\mathcal{F}$:

$$d_{IPM}(\mathbb{P},\mathbb{Q})=\sup_{f\in\mathcal{F}}|\mathbb{E}_{\mathbb{P}}[f(x)]-\mathbb{E}_{\mathbb{Q}}[f(y)]|,$$

with different IPMs resulting from the choice of F including the Wasserstein Distance, Maximum Mean Discrepancy and Total Variation Distance. Unlike the f-divergences, IPMs have the advantage that they are able to reward closeness of outcomes rather than only the ratio between probability of outcomes (Bellemare et al., 2017). The 1-Wasserstein distance d_W, hereafter referred to as the Wasserstein distance, can be defined in terms of the inverse distribution functions $F_{\mathbb{P}}^{-1}$ and $F_{\mathbb{Q}}^{-1}$:

$$\begin{aligned}d_W(\mathbb{P},\mathbb{Q})&=\sup_{f\in F_L}|\mathbb{E}_{\mathbb{P}}[f(x)]-\mathbb{E}_{\mathbb{Q}}[f(y)]|\\&=\int_0^1|F_{\mathbb{P}}^{-1}(u)-F_{\mathbb{Q}}^{-1}(u)|\,du.\end{aligned}$$

The Wasserstein distance, also referred to as the Earth-mover's distance, is central to the field of optimal transport (Villani et al., 2009), where it is conceptualised as the minimum cost of transporting mass from distribution $\mathbb{P}$ to obtain distribution $\mathbb{Q}$. It enables the comparison of distributions of different shapes and supports and the intercomparison of continuous and discrete distributions. However, Thorarinsdottir et al. (2013) note that it is also not a score divergence – or equivalently in the terminology of machine learning, Bellemare et al. (2017) point out that it has biased sample gradients, such that when minimising the distance using stochastic gradient descent, it may not converge or may converge to the wrong minimum.

Ghil (2016) first proposed the use of the Wasserstein distance alongside standard distributional moments in a climate modelling context, analysing parameter dependence of thermocline depth in an El Niño–Southern Oscillation (ENSO) model. Vissio et al. (2020) applied the Wasserstein distance as a more rigorous approach to global and regional ranking of CMIP6 model simulations of surface temperature, precipitation and sea ice cover.

B.1.3. | Integrated quadratic distance

The Integrated Quadratic (IQ) distance is calculated based on the squared distance between cumulative distribution functions $F_{\mathbb{P}}$ and $F_{\mathbb{Q}}$ as

$$d_{IQ}(\mathbb{P},\mathbb{Q})=\int_{-\infty}^{\infty}(F_P(x)-F_Q(x))^2dx.$$

Thorarinsdottir et al. (2013) propose the IQ distance for evaluation of climate simulations on the basis that it is a score divergence obeying proper scoring rules, demonstrating its relationship with the proper Continuous Ranked Probability Score used to measure probabilistic forecasts of a ground-truth scalar observation. It is similarly proposed[8] by Bellemare et al. (2017) as a metric that combines the advantages and overcomes the limitations of the Wasserstein distance and the KL divergence in that it has unbiased sample gradients and can also account for closeness between outcomes.

Thorarinsdottir et al. (2020) apply the IQ distance to evaluate North American and European extreme heat indices derived from CMIP5 and CMIP6 surface air temperature simulations. Vrac and Friederichs (2015) and Yuan et al. (2019) also use the IQ distance to evaluate the performance of novel bias correction and downscaling methods.

B.2 | Moments

The above divergence measures are used to quantify the distance between simulated and reanalysis distributions.

For comparison and to aid interpretation of these measures, it is useful to simultaneously characterise the location, dispersion and shape of the distributions being compared using some summary statistics. Conventional moments have some well-known limitations including oversensitivity to outliers, poor sampling efficiency and unsuitability for non-normal distributions (Wallis et al., 1974). Hosking (1990) proposed the alternative *L*-moments based on linear combinations of order statistics. Subsequent studies have demonstrated their superior empirical performance and sampling behaviour (Royston, 1992; Sankarasubramanian & Srinivasan, 1999). They are central to extreme value analysis (Silva Lomba et al., 2020) and are a standard approach for shape characterisation and parameter estimation of climatological distributions (Abdelmoaty et al., 2021; Papalexiou et al., 2020; Simolo et al., 2010).

APPENDIX C: CODE AND INTERACTIVE TOOL

Code and instructions to download data and reproduce the results presented in this paper can be accessed at: https://github.com/mvirdee/divergence_metrics. The interactive tool in Section 4.3 can be accessed in a Google Collaboratory notebook at: https://colab.research.google.com/drive/1DazaztOwLCdaIvonP944UEULc11yha-T.