



Sharper constraints without overconfidence: a scoring rule for perfect-model tests and climate projection weighting

Vittorio Del Tatto¹, Andrei Gavrilov¹, and Gustau Camps-Valls¹

¹Image Processing Laboratory, Universitat de València, València, Spain

Correspondence: Vittorio Del Tatto (vittorio.tatto@uv.es) and Gustau Camps-Valls (gustau.camps@uv.es)

Abstract. Projections of future climate carry large uncertainties, commonly quantified through the spread of a multi-model ensemble. The models contributing to such ensembles, however, are neither equally plausible nor independent. Weighting schemes address this by rewarding models that better reproduce the observed climate and penalizing redundant models, with both contributions computed from distances in a chosen predictor space. The performance weighting is governed by a free parameter, calibrated through a perfect-model test in which each model is treated in turn as a pseudo-reality. This parameter is set by following the philosophy of “constraining as much as possible without being overconfident”, monitoring how often the perfect models fall inside a pre-chosen confidence interval. Two problems follow: the test is not sensitive to the position of the pseudo-truth within the interval, and it assumes that some value of the performance parameter hits the target frequency, which is not guaranteed. Here, we reformulate perfect-model tests using the continuous ranked probability score (CRPS), a statistic that compares the entire weighted distribution with the pseudo-reality, and we select the performance parameter at its minimum. The use of this metric shifts the objective of weighting from constraining the projections to predicting well, with the narrowing of the ranges following as a consequence. On synthetic data, the resulting scheme constrains more than its coverage-based counterpart without suffering from overconfidence, provided that the model families are conservatively specified. We apply the two criteria to CMIP6 precipitation projections over land under SSP5-8.5, comparing predictor distances based on sea-level pressure fields and causal networks inferred from sea-level pressure time series. The CRPS-based weights narrow the very likely range by up to 32% for 2041–2060, while the end-of-century projections remain weakly constrained. Causal-network distances carry usable information and yield appreciable spread reductions, but they rest on shallower optima than a simple area-weighted RMSE, which remains the most reliable predictor distance we tested.

1 Introduction

Reliable climate projections on multi-decadal time scales are essential for climate risk assessment, which guides the adaptation and mitigation decisions taken by governments and planners. They inform estimates of future warming, shifts in precipitation, and changes in extremes that translate directly into impacts on water resources, agriculture, infrastructure, and human health. Because the future climate cannot be observed, these projections rest on numerical models. The Coupled Model Intercomparison Project (CMIP), currently in its sixth phase (CMIP6; Eyring et al., 2016), was developed by the World Climate Research Programme to coordinate the design and distribution of standardized climate model simulations of the past, present, and future,



so that results can be compared on a common footing across modeling groups and successive generations of models. Reducing the residual uncertainty of such ensembles, whether by improving the models (Eyring et al., 2024a, b) or by post-processing their output (Knutti et al., 2017; Brunner et al., 2020), is an active frontier in which statistical and machine-learning methods now play a central role.

30 The reliability of CMIP6 projections for a given climate variable depends not only on a best estimate, but on a faithful quantification of the associated uncertainty, which can be decomposed into three main sources. First, climate change depends on future anthropogenic emissions which are unknown, as they depend in turn on human decisions and future socioeconomic developments that can hardly be predicted. Second, each climate model features an internal variability which reflects the chaotic nature of climate dynamics: slightly different initial conditions in the Earth system's state develop into future scenarios which
35 become substantially different after a few years to at best a decade (Deser et al., 2012a). Third, different climate models represent different approximations of the real Earth system; they can differ in spatial resolution, geometry, the physical parameters entering the governing equations, and other modeling choices. This last source is mainly rooted in our limited knowledge of the Earth system, what Abramowitz et al. (2019) call *epistemic* uncertainty. Moreover, it also arises from the approximation schemes that each modeling group adopts to implement a shared understanding of the climate system with finite computational
40 resources.

The first source of uncertainty is handled by conditioning on a future scenario rather than predicting it: in CMIP6 models, future simulations are carried out under alternative Shared Socioeconomic Pathways (SSPs) coupled with Representative Concentration Pathways (RCPs). Together, they specify the expected human emissions under a future socioeconomic scenario (identified by an assumed population growth, economic and technological development, energy and land use, for example), and
45 the resulting radiative forcings, which directly enter the governing equations of the model. Once a coupled SSP-RCP scenario is specified, only the two remaining sources of uncertainty are left. The second contribution, the internal variability of the climate system, can be quantified by simulating the Earth system under a fixed scenario from different initial conditions. Still, it cannot be reduced: since it stems from the chaotic dynamics of the climate system, it would persist even in a model that perfectly reproduced reality (Deser et al., 2012a). In contrast, the spread of projections across models reflects our incomplete
50 knowledge of the climate system rather than an irreducible property of the system itself. It is therefore, in principle, reducible. This is the aim of an established line of research on projection weighting (Tebaldi and Knutti, 2007; Sanderson et al., 2017; Brunner et al., 2019; Merrifield et al., 2020).

Projection weighting was introduced for multi-model climate ensembles as an alternative to equal weighting (sometimes referred to as “model democracy”), the conventional scheme in which every simulation is given equal weight, the ensemble mean is taken as the best estimate, and the uncertainty is derived from the ensemble spread. As a first drawback, equal weighting
55 ignores that some models reproduce specific climate properties better than others. To account for this, projections can be weighted according to a performance criterion, rewarding simulations that agree more closely with observations (or reanalysis data). A second limitation is that equal weighting treats the ensemble projections as if they were independent realizations, whereas in practice they are not. Models share components, parameterization schemes, and even portions of code through their
60 common development history so that different models can produce similar projections for the same underlying reasons. A



further, more explicit source of dependence arises when an ensemble contains several simulations of the same model that differ only in their initial conditions or in the choice of physical parameters. These replicates are referred to as “members” of a given CMIP6 model. Under equal weighting, an over-represented model (or a family of “similar” models) pulls the best estimate towards its own behavior and narrows the apparent uncertainty, giving a false impression of agreement that reflects redundancy rather than independent corroboration. To penalize redundancy, a model can be down-weighted in proportion to its similarity to the rest of the ensemble, preventing clusters of near-identical simulations from dominating the result. The weighting scheme that accounts for both independence and performance was introduced by Knutti et al. (2017) and is sometimes referred to as ClimWIP (Brunner et al., 2020). Its application depends on the choice of an appropriate distance to compare ensemble members with the observational reference and among themselves. Alternatives to weighting exist, such as selecting a representative subset of the ensemble (Sanderson et al., 2015; Herger et al., 2018); here we work within the weighting framework, which retains all members and modifies only their contributions. It is worth recalling from the outset that weighting is not free of risk: an aggressive scheme can degrade a projection relative to equal weighting when the predictor carries little transferable information (Weigel et al., 2010). Guarding against this is precisely the purpose of the calibration procedure we revisit here.

Putting these ideas into practice raises the questions of *how much* performance should be rewarded and *how much* redundancy should be penalized. In ClimWIP, these two contributions are controlled by a pair of parameters, σ_D for performance and σ_S for independence, which set the distance scales over which a model is considered close to the observations or to another model. The choice of these parameters is not innocuous: it can change the weighted estimate and its uncertainty range substantially. At one extreme, making the performance weighting too aggressive concentrates almost all the weight on the single member closest to the historical observations, collapsing the ensemble to one projection and yielding overconfident results; at the other, very weak weighting recovers model democracy. A robust calibration must therefore sit between these limits. In the combined scheme, the two parameters are usually fixed sequentially. The independence parameter σ_S is set first, using prior knowledge of model dependence. The performance parameter σ_D is then determined through a *perfect-model test* (PMT): each model in turn is removed from the ensemble and treated as a pseudo-observation, the remaining models are weighted using their historical period as reference, and their weighted projections are compared to the perfect model’s future. The quantity traditionally monitored in a perfect-model test is the empirical coverage of a pre-chosen confidence interval (e.g., 90%), namely the fraction of perfect models’ projections which fall inside a predicted interval. The optimal performance parameter σ_D is selected as the smallest value for which the empirical coverage does not drop below its nominal level. This approach targets the most constraining weights that do not produce overconfidence.

The traditional implementation of perfect-model tests has a main drawback: it reduces the comparison between the left-out projection and the distribution of weighted projections to a single binary in-or-out decision, discarding information about the distribution’s shape. As a consequence, it is insensitive to where the truth lies within the interval, and does not provide any explicit control on the weighted distribution’s bias. A further problem is that the rule “the smallest σ_D whose coverage is not smaller than the required confidence” may not be satisfied for any σ_D (see, for example, the Supplement of Brunner et al., 2019). In this work, we propose a revised perfect-model test which accounts for these limitations. Instead of monitoring the coverage of a pre-chosen confidence interval, we compute an objective function, the continuous ranked probability score



(CRPS) between the weighted ensemble and the perfect model, and we select the performance parameter at its minimum. The CRPS compares the perfect model against the full weighted distribution and is sensitive to its bias, rewarding weighted projections that are concentrated around the reference rather than simply wide enough to contain it. Moreover, the CRPS is continuous in σ_D and finite at both of its limits, so a minimum is always attained. The use of the CRPS in climate projection weighting is not itself new (Brunner et al., 2019, 2020; O'Reilly et al., 2024), and it has been used elsewhere to select the strength of a Bayesian model-averaging weighting (Shin et al., 2020). Within ClimWIP, however, the CRPS has been used only downstream of the weighting, to evaluate the resulting distribution or select diagnostic sets, while σ_D itself continues to be calibrated on coverage.

Using synthetic data experiments, we show that when the distances carry sufficient information about the projections, the CRPS-based approach constrains more strongly than its coverage-based counterpart while predicting the reference more accurately. We also show, in the same setting, that a perfect-model test which ignores the model families (the groups within which redundancy is assumed to reside) is driven towards overconfident weights. Our results highlight the importance of careful model family selection and suggest that more conservative choices should be preferred to avoid overconfidence. We summarize each perfect-model test with a skill score, defined as the relative reduction in CRPS at the selected σ_D with respect to the independence-only limit ($\sigma_D \rightarrow +\infty$). This score informs about the suitability of the distance metric employed in ClimWIP and the reliability of the PMT-based weights.

We apply our revised perfect-model test to weight global CMIP6 precipitation projections over land under the SSP5-8.5 pathway over two separate time spans (2041-2060 and 2081-2100). Precipitation is among the climate variables whose future projections are most uncertain, particularly over the vulnerable and populated land areas that matter most for adaptation (Debeire et al., 2025). This uncertainty has several origins. Both thermodynamic and dynamic mechanisms govern the response of precipitation to warming: while the thermodynamic component is relatively well constrained by the Clausius–Clapeyron relation, the dynamic component, i.e. the response of the atmospheric circulation, is highly model-dependent and represents a dominant source of inter-model spread (Shepherd, 2014). Regional precipitation trends are also strongly affected by internal variability, which is largely mediated by large-scale atmospheric circulation and does not average out even on multi-decadal timescales (Deser et al., 2012b). As a result, the spread of precipitation projections across CMIP6 models remains wide, and reducing it is a central goal of model weighting (O'Reilly et al., 2024).

Following Debeire et al. (2025), we constrain precipitation projections using a predictor based on sea-level pressure (SLP), which characterizes the large-scale atmospheric circulation. The rationale is that both sources of spread described above run through the circulation, so that a model's fidelity to the observed SLP field should carry information about its precipitation response, as reported for CMIP5 (Nowack et al., 2020) and exploited for CMIP6 (Debeire et al., 2025). We exploit this known link using a conventional field-based distance (an area-weighted RMSE of the SLP climatology) and distances derived from causal-network representations of SLP dynamics, including the F_1 -score distance used in Debeire et al. (2025), the Structural Hamming Distance (SHD; Tsamardinos et al., 2006) and the structural intervention distance (SID; Peters and Bühlmann, 2015). The causal graphs that we employ were inferred by Debeire et al. (2025) using tools from causal discovery for time series (Runge et al., 2019b; Runge et al., 2023; Camps-Valls et al., 2023). Representing a climate model by a causal graph rather



than by a climatological field is attractive because the graph encodes the dynamical dependencies among circulation modes rather than their time-mean state, and is therefore a candidate for a process-level rather than a state-level comparison (Nowack et al., 2020; Runge et al., 2019a). Whether that promise translates into a more reliable constraint is one of the questions the present test is designed to answer. We find that the CRPS criterion selects systematically stronger weighting than coverage
135 does. The very likely range of the 2041–2060 change narrows by up to 32%, while the end-of-century range is constrained far more weakly. The causal-network distances do carry information (one of them yields the largest end-of-century reduction we obtain) but the perfect-model skill score places them on shallower optima than the area-weighted RMSE, which remains the most reliable distance in both periods despite being the simplest.

The rest of the paper is structured as follows. In Section 2, we first review ClimWIP and the standard approaches for tuning
140 the weight parameters; then, we introduce the CRPS-based implementation of the perfect-model test for tuning the performance parameter. In Section 3, we introduce an idealized model which relates distances between members in a space of “historical embeddings” to synthetic projection values. We exploit this model to study the CRPS-based weights across several regimes, accounting for varying levels of dependence between distances and projections, redundancy, and misspecification of the model families. In Section 4, we apply the introduced approach to CMIP6 projections; finally, we draw our conclusions in Section 5.

145 2 Methods

First, we introduce the notation used throughout this paper. We consider a multi-model ensemble composed of N members (replicas with different initial conditions or different physical parameters of the same climate model), possibly including a different number of members per model. We call $\mathbf{x}_i$ a representation of member i , built only from its historical realization, and we will refer to it as the “historical embedding” or “predictor” associated with member i . Another common term in
150 the ClimWIP literature is “diagnostic”. We name p_i a future projection associated with the same member under the SSP-RCP pathway of interest. We will assume that p_i is a one-dimensional statistic condensing information of a target climate variable in a region and time span of interest (in our application, the mean anomaly of precipitation over land in 2041–2060 or 2081–2100, in mm d^{-1}). Typically, p_i represents a variable which carries a marked ensemble uncertainty, in the sense that the distribution of the ensemble projections is too broad to inform decisions or risk assessment. In contrast, the “historical embedding” $\mathbf{x}_i$
155 might be a complex and high-dimensional representation (e.g., the spatial SLP field averaged over the historical period 1979–2014, or a causal graph encoding spatio-temporal relationships within components of the same field, inferred through causal discovery). The representation $\mathbf{x}_i$ is assumed to be informative about the member’s future projection p_i . Finally, we call $\mathbf{x}_{\text{ref}}$ the representation associated with observational or reanalysis products (here ERA5; Hersbach et al., 2020), and we call p_{ref} its unknown “true” projection. Members are grouped into *families* $\mu(i)$, the sets within which redundancy is assumed to reside;
160 the construction of this partition is discussed in Sects. 2.2 and 4.2.



2.1 Climate projection weighting by independence and performance – ClimWIP

The ClimWIP scheme introduced by Knutti et al. (2017) and Sanderson et al. (2015) assigns to member i the following weight:

$$w_i = C \frac{e^{-d(\mathbf{x}_i, \mathbf{x}_{\text{ref}})^2 / \sigma_D^2}}{1 + \sum_{j(\neq i)} e^{-d(\mathbf{x}_i, \mathbf{x}_j)^2 / \sigma_S^2}}, \quad (1)$$

165 where C is a constant (typically omitted for simplicity) fixed by requiring that the weights sum to 1: $\sum_{i=1}^N w_i = 1$. The argument of the exponentials, $d(\mathbf{x}_i, \mathbf{x}_{\text{ref}})$ and $d(\mathbf{x}_i, \mathbf{x}_j)$, denote distances between their historical embeddings. In ClimWIP applications, the two distances are often computed on different predictor fields; here, as in Debeire et al. (2025), both terms are computed from the same SLP representation. We use the term “distance” for $d(\cdot, \cdot)$ as a measure of dissimilarity, even when it does not satisfy all the axioms of a metric (notably the triangle inequality); we nonetheless always restrict to symmetric, or
 170 suitably symmetrized, distance functions.

The numerator of Eq. (1) is the performance contribution, which suppresses members whose distance $d(\mathbf{x}_i, \mathbf{x}_{\text{ref}})$ from the reference is large, according to a Gaussian law whose width is set by the performance parameter σ_D . In the limit $\sigma_D \rightarrow 0$, all weights vanish except the one associated with the $\mathbf{x}_i$ that minimizes $d(\mathbf{x}_i, \mathbf{x}_{\text{ref}})$, so the scheme collapses onto the single member closest to the reference. In the opposite limit $\sigma_D \rightarrow +\infty$, the performance contribution becomes uniform and is
 175 effectively switched off, and any distinction between weights is left to the denominator alone. The denominator of Eq. (1) implements the independence contribution: member i receives a penalty that grows with the number of embeddings $\mathbf{x}_j$ lying close to $\mathbf{x}_i$. The individual contributions to this penalty follow the same Gaussian law as the performance term, but with a width set by a distinct independence parameter σ_S . Despite this formal similarity, the limit $\sigma_S \rightarrow 0$ behaves very differently from the performance case. As $\sigma_S \rightarrow 0$, every term $e^{-d(\mathbf{x}_i, \mathbf{x}_j)^2 / \sigma_S^2}$ with $d(\mathbf{x}_i, \mathbf{x}_j) \neq 0$ vanishes¹, so the denominator reduces
 180 to 1 for every member: rather than selecting a single model, the independence contribution is switched off. In the opposite limit $\sigma_S \rightarrow +\infty$, every exponential tends to 1, so the denominator becomes the same constant (N) for all members. Therefore, the independence term discriminates between members only at intermediate values, where the Gaussian is wide enough to recognize clusters of similar members yet narrow enough to leave distinct ones unpenalized (Brunner et al., 2020). Both limits leave the weights uniform up to normalization, so the independence contribution is inert at either end of its range.

185 The weights in Eq. (1) strongly depend on the parameters σ_S and σ_D , which are typically tuned sequentially. First, the independence parameter σ_S is chosen as a distance scale which separates within-family and cross-family distances; then, the performance parameter is fixed using a perfect-model test.

2.2 Choice of the independence parameter σ_S

The independence parameter σ_S should ideally scale down an ensemble member by a factor that depends on the number of
 190 redundant members in the ensemble. However, neither the degree of redundancy nor the actual dependent entities are known a priori: as previously noted, redundancy is a feature of both members of the same CMIP6 model and members belonging

¹Excluding the pathological case in which two embeddings coincide exactly.



to models with a common development history or similar modeling choices. A popular strategy to set σ_S is to partition the ensemble into groups believed to possess a significant internal redundancy. Often, these groups are chosen as distinct CMIP6 models, so that their components are initial-condition members or perturbed-physics replicas of the same model. In contrast, motivated by the overconfidence pitfalls shown in Section 3, we adopt a more conservative approach, aggregating CMIP6 models with a known shared development history into the same “redundancy group”. We call such groups *model families*.

After defining the partition, σ_S can be chosen as a distance scale that optimally separates within-family and cross-family distances, penalizing pairs of members believed to be redundant while leaving as untouched as possible pairs that should be treated as independent. Here, we determine it by casting the separation as a binary classification between within-family and cross-family distances. For a candidate threshold, we compute the true-positive rate (TPR, the fraction of cross-family distances correctly placed above it) and the false-positive rate (FPR, the fraction of within-family distances wrongly placed above it). Then, we compute Youden’s J statistic (Youden, 1950), defined as the difference $J = \text{TPR} - \text{FPR}$, and we choose the threshold maximizing J . This criterion identifies the point of the receiver-operating characteristic (ROC) curve that lies farthest from the no-skill diagonal.

We highlight that alternative heuristics exist. For example, Merrifield et al. (2020) choose it from the distribution of within-family distances, fixing it two standard deviations below the mean of the same-model pairs. Brunner et al. (2020) calibrate σ_S from the typical distance between initial-condition members of a model, and Sanderson et al. (2017) fix it as a fraction of the distance between the best-performing model and the observations, calibrated so that the n variants of a model share a weight of about $1/n$.

2.3 Choice of the performance parameter σ_D via perfect-model test (PMT)

The performance parameter σ_D is set through a perfect-model test. The test is run after σ_S has been chosen, with the independence contribution held fixed. It evaluates a candidate σ_D against a pre-defined confidence interval (e.g. the 5th–95th percentile range) by iterating the following steps:

1. one ensemble member (the “perfect model”) is taken as a pseudo-reference;
2. weights are computed over the remaining members, discarding those belonging to the same “redundancy group” as the perfect model;
3. the required confidence interval is computed from the distribution of weighted projections;
4. the perfect model’s projection is compared to the confidence interval estimated at the previous step, recording whether it is inside or outside the range.

These steps are repeated by considering each member in turn as the perfect model; finally, the fraction of perfect models falling within the required confidence interval, which we will refer to as “PMT coverage”, is computed and compared to the nominal confidence level. An algorithmic presentation of this procedure, which we will refer to as “coverage-based PMT”, is reported in Appendix A (Algorithm A1).



Sweeping σ_D over a set of candidate values yields the PMT coverage as a function of the performance parameter. In the
 225 limit $\sigma_D \rightarrow 0$, the confidence range computed in any iteration of the perfect-model test reduces to a single projection value,
 associated to the closest member to the current perfect model. In this limit, excluding the possibility of identical projections,
 the perfect-model test yields zero PMT coverage. On the other hand, large values of σ_D lead to a plateau of the PMT coverage
 that should be only influenced by the independence term, and which should ideally be comparable with the required nominal
 confidence. To narrow the confidence intervals as much as possible without becoming overconfident, σ_D is commonly chosen
 230 as the smallest value for which the PMT coverage still equals or exceeds the nominal level (Knutti et al., 2017; Brunner et al.,
 2020).

The perfect-model test closely resembles a cross-validation procedure, the standard approach for hyperparameter tuning
 in machine learning: each ensemble member plays in turn the role of a held-out validation point, and the aggregated PMT
 coverage guides the choice of σ_D . In this perspective, the removal introduced in step 2 is essential. If the close relatives of
 235 the held-out member are left in the “training” set, the weighted projection distribution is artificially pushed towards the perfect
 model’s projection. This redundancy then inflates the PMT coverage, and the calibration is misled into accepting a smaller
 σ_D than it is actually safe. As a result, when the same σ_D is applied to the real observations, which have no duplicate within
 the ensemble, the resulting projection will be overconfident. Removing, at each iteration, all members belonging to the same
 model (or, more conservatively, the same model family) as the current pseudo-reference ensures that the held-out projection
 240 is compared only to reasonably independent projections (Abramowitz et al., 2019; Merrifield et al., 2020). We refer to this
 scheme as “leave-one-family-out” (LOF).

The success of a coverage-based PMT relies on the existence of at least one value of σ_D for which the PMT coverage
 is larger than or equal to the nominal coverage. However, this is not necessarily the case: the coverage attained by the test
 may remain below the nominal level for *every* σ_D , so that the prescription “the smallest σ_D that is not overconfident” has no
 245 solution (Brunner et al., 2020). The reference implementation of ClimWIP in the ESMValTool software treats the absence of
 such a value as a test failure (Andela et al., 2026). To circumvent this problem, some authors deliberately allow for a smaller
 PMT coverage than the required nominal coverage when the test fails (Brunner et al., 2019). A similar reduction can also be
 motivated by a preference for performance over independence in the weighting trade-off (Debeire et al., 2025). However, no
 quantitative criterion establishes how far below the nominal level one may go, and this remains a free choice of the analyst.
 250 In our implementation throughout this paper, we apply the coverage-based PMT conservatively: when a σ_D that satisfies the
 above requirement does not exist, we set it to the value that yields the largest PMT coverage. In case of ties, we select the
 smallest σ_D among the ties.

The coverage-based PMT has three further limitations, all following from the same source: it reduces the comparison be-
 tween a full (weighted) distribution and a reference projection into a single binary verdict. First, the verdict depends on a
 255 confidence range that must be fixed in advance and has no principled value. Second, within the interval, the test is blind to the
 location of the pseudo-reference. It therefore registers whether the weighted distribution is wide enough to contain the truth,
 but not whether it is centered on it, and it does not reward a probability mass concentrated near the truth. Third, being a fraction
 of successes over a finite number of iterations, the coverage can only assume a discrete set of values and thus behaves as a



step function of σ_D , so that the selected σ_D can jump discontinuously under small changes in the ensemble or in the σ_D grid.

260 All these limitations stem from discarding the distributional information available at each iteration, which can be avoided by replacing the coverage with a continuous score computed on the full weighted distribution.

2.4 A CRPS-based perfect-model test

In this section, we review the continuous ranked probability score (CRPS) and show how it can be used in the perfect-model test in place of the coverage metric.

265 In our context, the CRPS measures the agreement between a (weighted) projection distribution and a single reference projection, generalizing the mean absolute difference between a prediction and a target. Consider one PMT iteration, in which member k is left out and its projection $\hat{p} = p_k$ plays the role of the pseudo-reality. Let $p_1, \dots, p_M$ be the projections of the M surviving members and $w_1, \dots, w_M$ their weights, normalized so that $\sum_{i=1}^M w_i = 1$. Let

$$F(z) = \sum_{i=1}^M w_i \mathbb{1}\{p_i \leq z\} \quad (2)$$

270 be the CDF of the weighted distribution that assigns mass w_i to each projection p_i , and $H_{\hat{p}}(z) = \mathbb{1}\{z \geq \hat{p}\}$ the CDF of the reference projection (a Heaviside step function jumping from 0 to 1 at $z = \hat{p}$). The CRPS can be written as the integrated squared difference between the two CDFs,

$$\text{CRPS}(F, \hat{p}) = \int_{-\infty}^{+\infty} [F(z) - H_{\hat{p}}(z)]^2 dz. \quad (3)$$

Therefore, the CRPS measures the difference between the entire predictive distribution, represented by its empirical CDF, and 275 the step-function CDF of the reference, by accumulating the mismatch across all thresholds z rather than at a single quantile or interval. It reduces to the absolute error when F is a point mass, and it is reported in the units of the projection variable.

For the weighted empirical measure of Eq. (2), Eq. (3) admits the equivalent *energy form*

$$\text{CRPS}(\{p_i, w_i\}, \hat{p}) = \sum_{i=1}^M w_i |p_i - \hat{p}| - \frac{1}{2} \sum_{i,j=1}^M w_i w_j |p_i - p_j|. \quad (4)$$

The first term of Eq. (4) rewards proximity of the ensemble members to the reference projection. In contrast, the second 280 quantifies the internal spread of the predictive distribution: the CRPS therefore decreases when the members are distributed around the reference, but penalizes a distribution that becomes arbitrarily narrow away from it. Whereas coverage only records whether the reference falls inside a pre-chosen interval, the CRPS is sensitive both to the location of the reference relative to the bulk of the distribution and to the dispersion of the distribution itself, and it varies smoothly with σ_D .

A CRPS-based PMT is obtained by replacing steps 3 and 4 of Section 2.3 with the computation of the CRPS between the 285 weighted projection distribution and the pseudo-reference, following Eq. (4). Averaging the resulting CRPS values over all PMT iterations yields a single score for each candidate σ_D (Algorithm A2, Appendix A), and the performance parameter is



selected as the value that minimizes this average,

$$\langle \text{CRPS}(\sigma_D) \rangle = \frac{1}{N} \sum_{k=1}^N \text{CRPS}(F_{\sigma_D}^{(k)}, p_k), \quad (5)$$

$$\sigma_D^* = \arg \min_{\sigma_D} \langle \text{CRPS}(\sigma_D) \rangle, \quad (6)$$

where $F_{\sigma_D}^{(k)}$ is the weighted distribution of the survivors in iteration k . In practice the minimization is a grid search: we evaluate Eq. (5) on a logarithmically spaced grid of σ_D spanning three decades around the typical scale of the member–reference distances, and we retain the global minimum over the grid. The objective is not guaranteed to be unimodal in σ_D , so a local search could converge to a spurious optimum; with a one-dimensional parameter, exhaustive evaluation is cheap and we prefer it.

3 Synthetic data experiments

To illustrate our revised perfect-model test and compare it with its coverage-based counterpart, in this section we construct a simple model that links a predictive space $\mathbf{x}$, where member-member and member-reference distances are computed, to synthetic random variables that serve as projections. With this model, we compare the weighted projections from coverage- and CRPS-based PMTs, identify their failure modes in a controlled setting, and define a skill score that reports the reliability of the selected performance parameter.

To mimic the presence of model families with redundant information, we construct the $\mathbf{x}$ space by extracting points (representing ensemble members) from different Gaussian clusters (representing model families) in a three-dimensional Euclidean space (Fig. 1a). We regard the representation $\mathbf{x}_i$ associated with member i as its *historical* state, for which we will use the notation $\mathbf{x}_i = \mathbf{h}_i(t_1)$. The rationale behind the cluster structure is that members belonging to the same family should lie closer, on average, than members of different families (Figs. 1b). The definition of the Gaussian mixture is given in Appendix B.

Given a member with representation $\mathbf{x}_i$, we further assign it a synthetic projection p_i according to the model described in the following. First, we generate a *future* state $\mathbf{h}_i(t_2)$ through a stationary random process emulating the drift of a climate state over the projection horizon. Each member’s state has two components: the centroid of its family, shared among all members of that family, and a member-specific fluctuation around it. We model both with the same autocorrelation $\rho \in [0, 1]$ (see Eqs. (B2) and (B3)), so that a single parameter governs the joint decorrelation of the ensemble between t_1 and t_2 . Varying ρ is then equivalent to varying the length of the projection horizon relative to the autocorrelation time of the process: large ρ corresponds to a short projection horizon, over which the future state remains close to the historical one; small ρ corresponds to a long projection horizon, over which the future state has largely decorrelated from the historical one. Equivalently, ρ measures how much memory of the historical state survives into the projection, and we will use the two readings interchangeably.

Then, we define the scalar projection p_i as a linear projection of the future state $\mathbf{h}_i(t_2)$ (see Eq. (B4)). Fig. 1c reports the projection distributions in an illustrative example. Finally, we complement the model with a reference pair $(\mathbf{x}_{\text{ref}}, p_{\text{ref}})$, playing the role of the observational or reanalysis product. The reference is a single member of its own cluster and its projection is

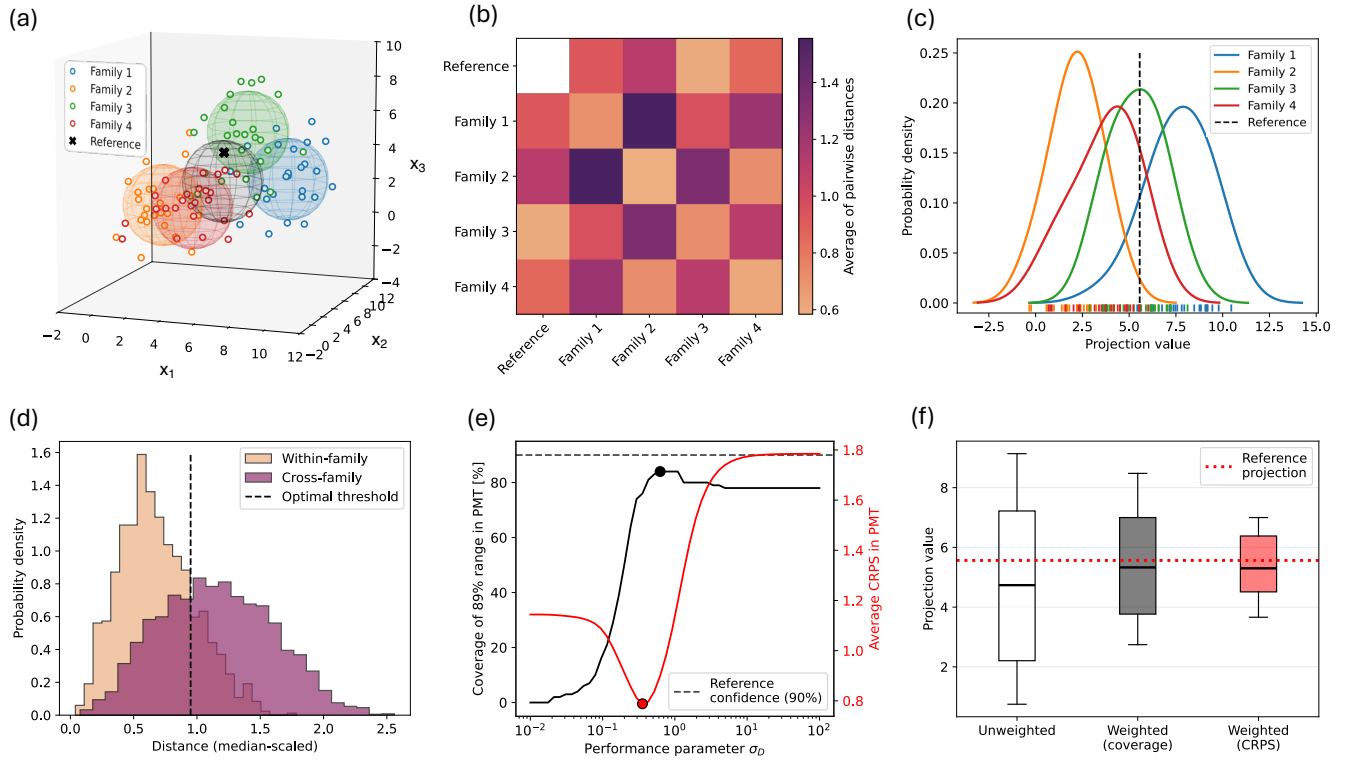


Figure 1. Illustration of synthetic data model and weight construction via perfect-model test. **(a)** Depiction of space $\mathbf{x}$: clusters represent ensemble families, points are family members. A reference point, playing the role of an observational or reanalysis product, is extracted from its own cluster. **(b)** Inter-family distance matrix, computed over the points of panel **(a)** using the Euclidean distance. Each entry reports the average member-member distance for all member pairs in the corresponding families. The entry (reference, reference) is empty, as a single synthetic reference is extracted. **(c)** Projection distributions corresponding to the points of panel **(a)**, generated using $\rho = 0.85$ in Eq. (B5). The reference projection is shown as a black dashed line. **(d)** Selection of an optimal threshold separating within-family and cross-family distances, which can be used to set the independence parameter σ_S . **(e)** Results of the perfect-model test as a function of the performance parameter σ_D , monitoring the coverage of the 90% confidence interval (black curve) or the average CRPS across PMT iterations (red curve). The selected values are marked by dots; note that the CRPS minimum lies at a smaller σ_D than the coverage maximum. **(f)** Projection distributions under the different weighting schemes. The center line is the mean, the box spans the 17th–83rd percentile range, and the whiskers span the 5th–95th percentile range. The reference projection is the red dotted line.

obtained through the same linear readout. Its construction rests on two assumptions: that $\mathbf{x}_{\text{ref}}$ is not an outlier in predictor space, and that the multi-model ensemble does not develop a systematic drift relative to the reference. The first assumption, which we
 320 verify on the real CMIP6 dataset (Section 4.3), is implemented in our model by placing the reference cluster at the barycenter of the family centroids at t_1 . The second assumption is imposed by assigning to the reference cluster the ensemble-mean drift, so that it remains at the barycenter of the family centroids at t_2 as well. Unlike the first, this assumption cannot be verified on a real multi-model ensemble, since it concerns the future evolution of the reference relative to the models.



Two parameters govern the regimes of interest. The autocorrelation ρ tunes the predictability of the projection from the representation: at $\rho = 1$, p_i is a deterministic function of $\mathbf{x}_i$, the strongest link that the weighting can exploit; at $\rho = 0$, the future centroid positions are independent of their historical values, and the historical representation carries no information about projections across families. The only residual predictability is within-family: members of the same family share the same future centroid draw and therefore remain mutually correlated, even though that shared draw is unpredictable from the historical state. This is precisely the non-transferable signal that the leave-one-family-out scheme is designed to discard (Section 2.3).

The cluster spread ξ , relative to the separation between the family centroids, sets the overlap between within-family and cross-family distances; we choose it so that this separation is comparable to the one observed for the CMIP6 ensemble (Section 4). The generative equations, the cluster geometry and the exact parameter values employed are reported in Appendix B.

Using this model, we construct ClimWIP weights by first selecting the independence parameter σ_S (Fig. 1d), following the optimal threshold criterion of Section 2.2. Distances are rescaled by their median, so that metrics with different units and ranges (Section 4) share a common scale and the same grid of σ_D values can be searched for all of them. Then, we apply the perfect-model test as a function of the performance parameter σ_D , by monitoring the PMT coverage of the 90% confidence interval (5th-95th percentile range) and the average CRPS across PMT iterations (Fig. 1e). Fig. 1e shows a typical failure scenario of the coverage-based PMT test: the coverage remains below the required reference confidence for all values of σ_D , so that the optimal performance parameter must be chosen according to an alternative criterion. When this scenario occurs, we select σ_D as the value yielding the largest PMT coverage; in case of ties, we select the smallest σ_D . The corresponding weighted projections are shown in the boxplots of Fig. 1f (lower panel), showing in both cases a reduced bias and a reduced spread with respect to the unweighted distribution, with a more marked constraint using the optimal σ_D from the CRPS-based scheme.

The example illustrated in Fig. 1 depends on a single realization of the historical embeddings $\{\mathbf{x}_i\}$ and a single extraction of the corresponding projections $\{p_i\}$. To carry out a systematic comparison of the two parameter-selection approaches, we study their behavior under repeated and independent drawings of $\{\mathbf{x}_i\}$ and $\{p_i\}$, as a function of the autocorrelation parameter ρ .

3.1 Systematic comparison of coverage-based and CRPS-based perfect-model tests

We carry out the comparison using three statistics, as shown in Fig. 2: uncertainty reduction, predictive skill improvement, and the reliability of the selected performance parameter σ_D . The uncertainty reduction achieved by weighting is computed as the relative variation of the 5th-95th percentile range (“ $W_{90\%}$ ”) from the unweighted to the weighted projection distributions, averaged over 100 independent extractions of $\{\mathbf{x}_i\}$ and $\{p_i\}$ (first row of Fig. 2):

$$\left\langle \frac{W_{90\%}^{\text{weighted}} - W_{90\%}^{\text{unweighted}}}{W_{90\%}^{\text{unweighted}}} \right\rangle, \quad (7)$$

where $\langle \cdot \rangle$ denotes the average over the independent extractions. In each resampling, the points $\{\mathbf{x}_i\}$ and their projections $\{p_i\}$ are regenerated with the same model parameters (Appendix B) using a different random seed, and the weighting pipeline previously described is re-run independently. The reference pair $(\mathbf{x}_{\text{ref}}, p_{\text{ref}})$ is also resampled independently at each extraction. A negative value of Eq. (7) denotes a reduction in the spread of the projection distribution, as an effect of weighting. As a

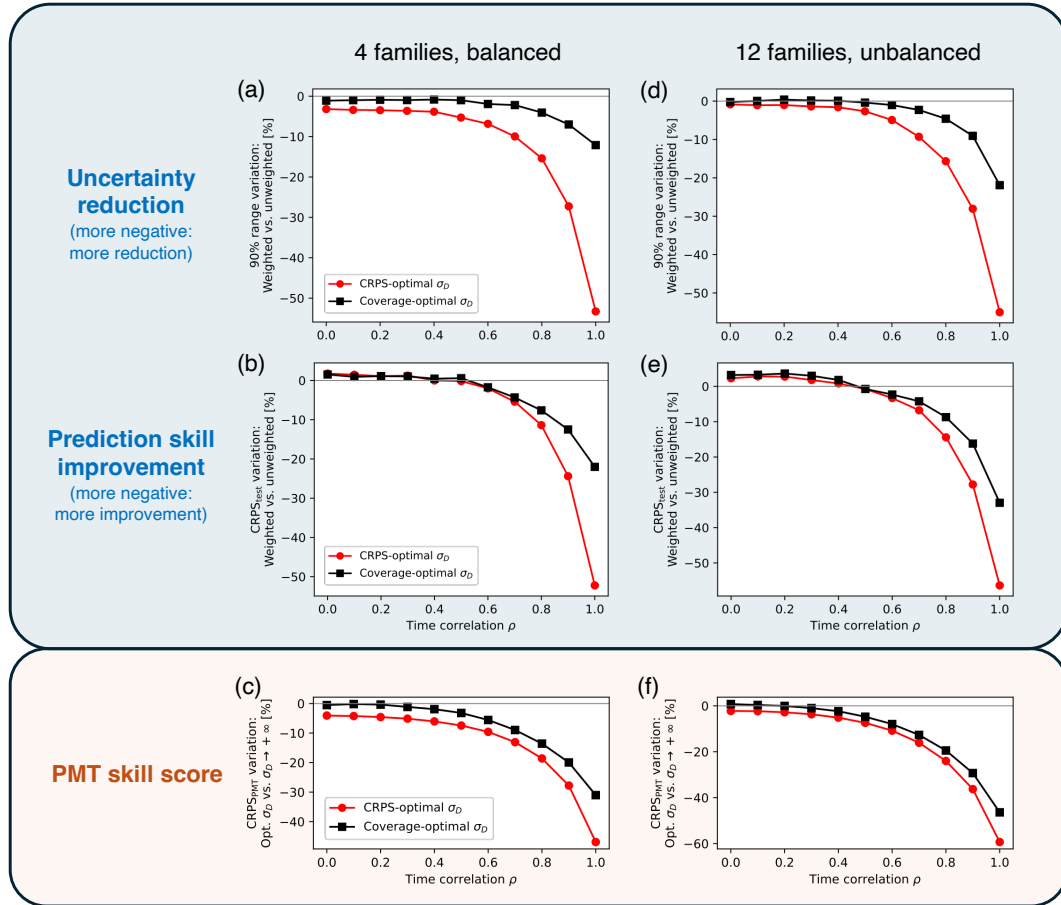


Figure 2. Systematic comparison of the optimal weightings selected through coverage-based (black curves) and CRPS-based (red curves) PMTs, as a function of the autocorrelation parameter ρ in the model. Panels (a)–(c) show results for a data-generation model with 5 families of 25 members each, panels (d)–(f) for 12 families with a variable number of members per family, mimicking the imbalance of the CMIP6 dataset (Table 1). (a) and (d): Relative variation of the 90% confidence range of the weighted distribution with respect to the unweighted distribution, see Eq. (7). (b) and (e): Relative variation of $\text{CRPS}_{\text{test}}$, see Eq. (8). (c) and (f): Relative variation of CRPS_{PMT} , see Eq. (9). In all panels, negative values are the desirable outcome. Each point averages 200 independent realizations of $\{\mathbf{x}_i\}$ and $\{p_i\}$.

second evaluation metric, we monitor the CRPS of the weighted and unweighted distributions with respect to the reference projection p_{ref} , which we denote by $\text{CRPS}_{\text{test}}$ to avoid confusion with the CRPS estimates within the perfect-model test. We then compute a relative variation (second row of Fig. 2) analogous to Eq. (7),

$$360 \left\langle \frac{\text{CRPS}_{\text{test}}^{\text{weighted}} - \text{CRPS}_{\text{test}}^{\text{unweighted}}}{\text{CRPS}_{\text{test}}^{\text{unweighted}}} \right\rangle. \quad (8)$$

While Eq. (7) quantifies the reduction in spread due to weighting, Eq. (8) measures the improvement in predictive skill relative to the reference, as captured by the CRPS (Section 2.4); negative values indicate that weighting brings the projection distribu-



tion “closer” to p_{ref} . Clearly, the variation in predictive skill cannot be computed for real multi-model ensembles, as p_{ref} is not accessible.

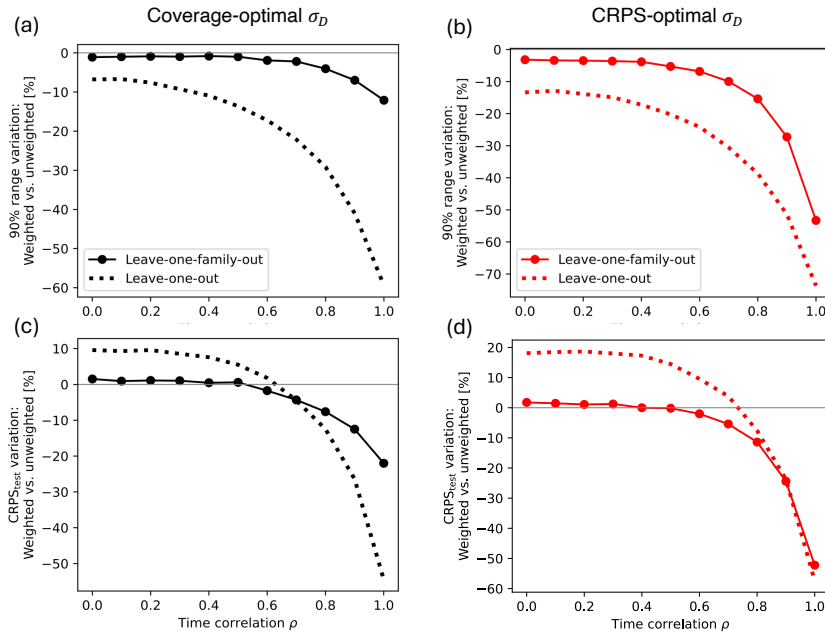


Figure 3. Comparison of leave-one-family-out and leave-one-out PMTs, as a function of the autocorrelation parameter ρ , for the 5-family model. **(a)-(b):** Relative variation of the 90% confidence range of the weighted distribution w.r.t. the unweighted distribution, for coverage-based and CRPS-based weight constructions, respectively. **(c)-(d):** Relative variation of $\text{CRPS}_{\text{test}}$ in the same cases.

365 We complement our evaluation metrics with a measure of reliability of the σ_D -selection scheme (third row of Fig. 2). For the optimal σ_D selected by each scheme, we compute the CRPS between the weighted distribution of the left-out members and the perfect model’s projection, averaged over all PMT iterations. We denote this quantity, which is exactly the minimized objective function within the CRPS-based scheme, by $\text{CRPS}_{\text{PMT}}(\sigma_{D,\text{opt}})$. Then, we compare it with the same quantity in the absence of the performance weighting ($\sigma_D \rightarrow +\infty$), by computing the relative difference

$$370 \left\langle \frac{\text{CRPS}_{\text{PMT}}(\sigma_{D,\text{opt}}) - \text{CRPS}_{\text{PMT}}(\sigma_D \rightarrow +\infty)}{\text{CRPS}_{\text{PMT}}(\sigma_D \rightarrow +\infty)} \right\rangle. \quad (9)$$

The average is over independent extractions of $\{\mathbf{x}_i\}$ and $\{p_i\}$ as in the previous evaluation metrics. This quantity plays different roles in the two schemes. For the CRPS-based scheme, the relative difference in Eq. (9) is, by construction, the relative depth of the minimum from which σ_D is selected, and thus a direct measure of how well-determined that minimum is. A pronounced (negative) value indicates that the selected σ_D substantially reduces the perfect-model CRPS; a value close to zero indicates that performance weighting adds little to the independence-only weighting, so that σ_D is poorly constrained and the resulting weights less reliable. For the coverage-based scheme, σ_D is selected without reference to this quantity, so Eq. (9) does not diagnose the reliability of that selection. Nonetheless, we report it for the coverage-optimal σ_D as a post-selection check to



assess whether the weighting improves the projection distribution in the perfect-model test. Evaluating a coverage-selected weighting through the perfect-model CRPS follows Brunner et al. (2019).

380 We use the above metrics to compare the two σ_D -selection schemes in two scenarios: a balanced multi-model ensemble of 5 families with 25 members each (left panels of Fig. 2), and an unbalanced ensemble of 12 families and 95 total members, mimicking the imbalance of the CMIP6 dataset used later (Table 1; right panels of Fig. 2). In both cases we study the results as a function of the autocorrelation parameter ρ . Figures. 2a and d show that, in the strong-memory limit ($\rho \rightarrow 1$), the spread of the projection distribution can be reduced by both coverage-based and CRPS-based weights (red and black curves, respectively),
 385 with the second providing a more marked reduction. On the other hand, in the weak-memory limit ($\rho \rightarrow 0$) neither measure can, on average, provide any significant uncertainty reduction. However, a greater reduction in spread does not imply a better projection distribution: taken on its own, the narrower range produced by the CRPS-based σ_D could reflect overconfidence. To exclude this, in Figs. 2b and e we show the relative variation of $\text{CRPS}_{\text{test}}$, accounting for both spread and bias of the final distribution. In the strong-memory regime, both weighting schemes improve overall prediction quality relative to the
 390 unweighted case, and the CRPS-based weights yield the best projection distribution. Therefore, the proposed scheme improves two desirable properties of the weighted distribution at once: a stronger reduction in spread and a better prediction skill.

The PMT skill score of Eq. (9) is shown in Figs. 2c and f. Consistent with the other metrics, CRPS_{PMT} becomes increasingly negative as ρ approaches 1: at high ρ , the predictor carries enough transferable signal for the CRPS-based perfect-model test to identify a clear optimum, while at low ρ , the objective flattens and σ_D becomes less identifiable. The variation is systematically
 395 more negative for the CRPS-based scheme than for the coverage-based one; this is expected, since the CRPS-based scheme selects σ_D precisely at the minimum of this objective, whereas the coverage criterion does not. Beyond comparing the two schemes, the skill score is most useful as a within-scheme measure, comparing the reliability of the optimal σ_D across different scenarios for a fixed σ_D -selection rule. We exploit this in Section 4, where this score is used to compare alternative distances $d(\cdot, \cdot)$ in ClimWIP and to assess, for each of them, whether the perfect-model test yields a reliable performance parameter.

400 The results shown so far are obtained by implementing the perfect-model test in a leave-one-family-out (LOF) scheme, where the families coincide with the Gaussian clusters. In Fig. 3, we show that neglecting this family structure and implementing the perfect-model test in a simple leave-one-out scheme leads to overconfidence. In panels a and b, we compute the variation of the 90% range achieved by the coverage-optimal and CRPS-optimal weights as a function of ρ . The solid lines represent the same curves shown in Fig. 2d; the dotted curves show the uncertainty reductions obtained by neglecting the family structure
 405 of the data and implementing the perfect-model tests with simple leave-one-out schemes. The reduction achieved in the latter case is more significant for both σ_D -selection rules. However, panels c and d show that in both cases these results are prone to overconfidence: when ρ is small enough ($\rho < 0.6$ in the coverage-based scheme, $\rho < 0.7$ in the CRPS-based), the variation of $\text{CRPS}_{\text{test}}$ becomes positive if families are neglected, denoting a decrease of the prediction skill compared to the unweighted case. Not surprisingly, this effect is more pronounced in the CRPS-based scheme: consistent with the previous results, the
 410 CRPS-based weights appear to impose a further constraint that, in this case, exploits the non-transferable relationship between the left-out member and its family-mates.



This confirms that a correct specification of the family structure is essential. There is, however, no ground-truth criterion to decide whether a given family specification is correct: the dependence between models is a matter of degree rather than a clear-cut partition, and any grouping is necessarily an approximation (Abramowitz et al., 2019; Annan and Hargreaves, 2017). For this reason, more conservative groupings should be preferred to avoid overconfidence. Prior knowledge of shared development history can be complemented with data-driven clustering of the model representations, which we exploit in Section 4 to inform and cross-check the family definition.

4 Application to CMIP6 projections

In this section, we apply the coverage-based and CRPS-based weighting schemes to a multi-model CMIP6 ensemble. The projections that we employ are global anomalies of precipitation over land, time-averaged over the years 2041–2060 and 2081–2100, under the global warming scenario SSP5-8.5, and expressed in mm d^{-1} . The baseline for computing the anomalies is the period 1995–2014. The representations $\mathbf{x}_i$ are built over sea-level pressure (SLP) historical simulations in the time span 1979–2014, while the reference $\mathbf{x}_{\text{ref}}$ is constructed over ERA5 SLP (Hersbach et al., 2020) in the same time range. The nature of these representations and the distances employed in our study are specified in Section 4.1.

Our ensemble comprises 154 members from 23 distinct CMIP6 models and coincides with the multi-model ensemble used in Debeire et al. (2025). Except for CanESM5, whose members belong to two distinct physics variants, the members of each model differ only in their initial conditions. While historical simulations are available for all such members, only a subset of them ($N = 95$) is associated with projections of precipitation under the SSP5-8.5 scenario, based on the data available on the ESGF (Cinquini et al., 2014) at the time of the analysis. To make use of all available data, the calculation of σ_S and the validation of the family partition were based on historical simulations for all members, including those without available projections. For the remaining steps, which involve estimating σ_D via a perfect-model test, we considered only members coupled to available projections. A complete list of the CMIP6 models and members entering each stage of the analysis (the SLP RMSE, the causal-graph distances, the SSP5-8.5 projections) is reported in Table S1 in the Supplement. Because the ensemble coincides with the one of Debeire et al. (2025), the sea-level pressure time series from which the causal graphs were obtained, as well as the precipitation projections over land, are taken directly from that study, to which we refer for the corresponding preprocessing details (and to Debeire (2025, code) for the implementation). The SLP fields entering the RMSE were instead produced for the present work with ESMValTool v2.14.0 (Righi et al., 2020), using the recipe provided with the data. Starting from the same daily SLP data over 1979–2014, the fields are remapped onto a regular $2.5^\circ \times 2.5^\circ$ grid between 87.5°S and 87.5°N with a conservative (area-weighted) scheme; they are then averaged to monthly means and detrended linearly at each grid cell. At variance with Debeire et al. (2025), who standardize the daily anomalies by the standard deviation of a long-term daily climatology to improve the stationarity required by the causal discovery step, we retain the physical units and the seasonal cycle of the fields, as required for the RMSE to measure a distance between simulated and observed seasonal cycles.

4.1 Embeddings and distances

445 In this application, we use the coverage-based and CRPS-based perfect-model tests for ClimWIP to compare how different distances $d(\mathbf{x}_i, \mathbf{x}_j)$ constrain the precipitation projections. To better account for the effect of seasonalities, each representation $\mathbf{x}_i$ is first decomposed in four representations $\mathbf{x}_i^s$, each computed on historical data restricted to a season s ($s \in \{\text{DJF}, \text{MAM}, \text{JJA}, \text{SON}\}$). For each metric that we test, $d(\mathbf{x}_i, \mathbf{x}_j)$ is computed as the average over the four seasonal contributions:

$$450 \quad d(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{4} \sum_s d(\mathbf{x}_i^s, \mathbf{x}_j^s). \quad (10)$$

Table 1. Definition of the CMIP6 model families used throughout this work.

Model family	Models	Number of projections
MetUM	ACCESS-CM2, ACCESS-ESM1-5, HadGEM3-GC31-LL, HadGEM3-GC31-MM, KACE-1-0-G, UKESM1-0-LL	21
BCC	BCC-CSM2-MR	1
CESM2	CESM2, CESM2-WACCM, NorESM2-LM	5
CNRM	CNRM-CM6-1, CNRM-ESM2-1	11
CanESM5	CanESM5	14
EC-Earth3	EC-Earth3, EC-Earth3-Veg	5
FGOALS	FGOALS-g3	2
INM	INM-CM5-0	1
IPSL	IPSL-CM6A-LR	5
MIROC	MIROC-ES2L, MIROC6	17
MPI	MPI-ESM1-2-HR, MPI-ESM1-2-LR	12
MRI	MRI-ESM2-0	1

First, we test a standard area-weighted RMSE between the spatial SLP fields of two members i and j . To compute it, we first extract a time-averaged SLP field for each season over the historical period (1979–2014), yielding a $2.5^\circ \times 2.5^\circ$ grid of SLP values for each season and ensemble member. We denote by $x_i^s(\phi, \lambda)$ the component of such a field at latitude ϕ and longitude λ , for member i and season s . The same construction is applied to SLP data from ERA5. Then, the area-weighted RMSE for each season is computed as

$$d(\mathbf{x}_i^s, \mathbf{x}_j^s)_{\text{RMSE}} = \left[\frac{\sum_{\phi\lambda} w_{\phi\lambda} (x_i^s(\phi, \lambda) - x_j^s(\phi, \lambda))^2}{\sum_{\phi\lambda} w_{\phi\lambda}} \right]^{\frac{1}{2}}, \quad (11)$$

where the sum runs over all (ϕ, λ) grid points and the weights are computed as $w_{\phi\lambda} = \cos(\phi)$.

The additional metrics that we test are based on the causal graph representations obtained by Debeire et al. (2025). In this case, the representations $\{\mathbf{x}_i^s\}$ (and $\mathbf{x}_{\text{ref}}^s$) are directed acyclic graphs (DAGs) representing the interactions between different

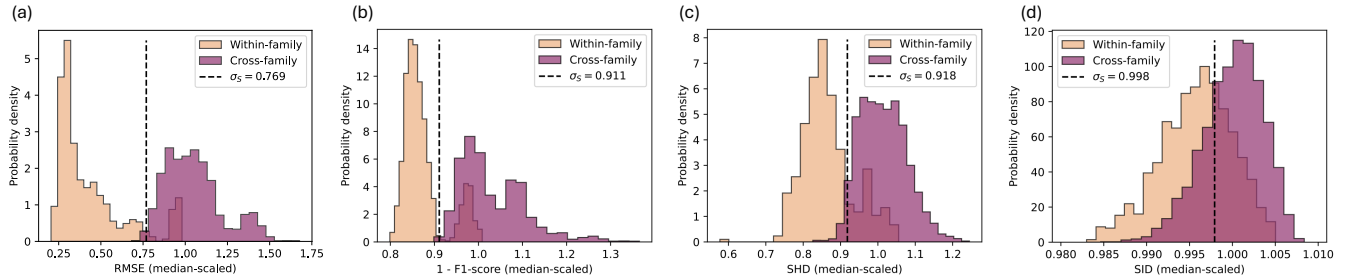


Figure 4. Selection of σ_S as a threshold that optimally splits within-family and cross-family distances, for (a) RMSE, (b) $1 - F_1$ -score, (c) SHD and (d) SID. The optimal thresholds are reported in the legends.

460 components of the SLP field through time, for each season. These graphs were inferred by Debeire et al. (2025) following a pipeline originally proposed by Nowack et al. (2020), which includes a sequence of preprocessing, dimensionality reduction, and causal discovery steps, using as input the SLP historical data of each member (and ERA5). We direct the reader to these references for details about each inference step.

In Debeire et al. (2025), these representations were used to build a graph distance used in the ClimWIP weighting scheme of
 465 Eq. (1). Specifically, the authors recast graph comparison as an edge-classification problem: the graph associated with member i is treated as a “predictor” of edges and the graph associated with member j as the “ground truth”. Then, the discrepancy between the two graphs can be computed through the F_1 -score in this binary classification (an edge can be either present or absent, since DAGs are unweighted graphs), which quantifies the agreement between the two edge sets as the harmonic mean of precision and recall. A measure with the basic property of a distance (more similar objects should have a smaller separation)
 470 can then be computed as $1 - F_1$. In the following, we test two alternative graph metrics: the structural Hamming distance (SHD, Tsamardinos et al. 2006), which counts the number of edges on which the two structures disagree, and the structural intervention distance (SID, Peters and Bühlmann 2015; Henckel et al. 2024), which exploits the causal interpretation of the graphs in terms of interventions: rather than counting edge mismatches, it counts the pairs of nodes for which the two graphs imply different interventional (causal) effects, quantifying how often using one graph in place of the other would yield an
 475 incorrect causal-effect estimate. The three dissimilarities therefore probe increasingly demanding notions of agreement, from set overlap (F_1 , SHD) to agreement of the implied causal effects (SID). Additional details on the definition and computation of the graph distances are reported in Appendix C.

4.2 Choice of model families

In the synthetic experiments, we stressed the importance of correctly specifying the model families. Their definition determines
 480 the choice of the independence parameter σ_S , and it significantly affects the outcome of the perfect-model test when applied via LOF, leading to an overconfident choice of the performance parameter σ_D when the families are misspecified.

Our definition of model families (Table 1) is based on knowledge of common development history and shared model components among distinct models. This choice reflects the fact that the LOF scheme is designed to remove dependencies known a



priori, and follows the practice of characterizing CMIP interdependence through model genealogy (Masson and Knutti, 2011; Knutti et al., 2013; Boé, 2018; Kuma et al., 2023). We group models that share a physical core, an atmospheric component, or a direct development lineage, since such models are known to produce correlated historical and future behavior (Knutti et al., 2013; Merrifield et al., 2020). The clearest example is the MetUM family, which collects all models built on the UK Met Office Unified Model atmosphere (Walters et al., 2019): the two HadGEM3-GC3.1 configurations, UKESM1-0-LL, KACE-1-0-G and ACCESS-CM2 (Bi et al., 2020) all use the MetUM-HadGEM3-GA7.1 atmosphere (Walters et al., 2019), while ACCESS-ESM1-5 (Ziehn et al., 2020) is built on an earlier, HadGEM2-era UM atmosphere of the same lineage. This grouping is independently supported by Merrifield et al. (2023), who identify essentially the same set as a single family from inter-member distances in surface-temperature and sea-level-pressure climatologies. The one difference is ACCESS-ESM1-5, which they assign to no family as characterized by an earlier-generation HadGAM2 atmosphere; our partition is more conservative in this respect. The remaining families follow the same principle. The CESM2-based models share the CESM2 framework and the CAM6 atmospheric lineage, though not an identical atmosphere: CESM2 uses the low-top CAM6, CESM2-WACCM the high-top WACCM6 with interactive middle-atmosphere chemistry (Danabasoglu et al., 2020), and NorESM2-LM a modified version, CAM6-Nor, with its own aerosol module and a different ocean (Seland et al., 2020). The pairs of models CNRM (Voldoire et al., 2019; Séférian et al., 2019), EC-Earth3 (Döscher et al., 2022), MIROC (Tatebe et al., 2019; Hajima et al., 2020), and MPI-ESM1-2 (Mauritsen et al., 2019; Müller et al., 2018) are each grouped by shared institutional development and common core components. Models with no clear relative in our ensemble (BCC, CanESM5, FGOALS, INM, IPSL, MRI) are treated as singleton families. We note that our grouping is an inevitable simplification: dependence in reality lies on a continuum rather than forming disjoint sets (Abramowitz et al., 2019; Annan and Hargreaves, 2017), and some assignments are stronger than others (for instance ACCESS-ESM1-5 uses an earlier-generation UM atmosphere than the other members of its family). Our aim is a conservative partition of the strongest known redundancies, which is what the leave-one-family-out scheme requires to avoid the overconfidence pitfalls discussed in Section 3. Since a coarser partition removes more members at each LOF iteration, it makes the test harder to pass and the selected weighting weaker. The bias introduced by an imperfect grouping is therefore towards under-constraining, which is the direction we prefer to err in.

The family split is validated against a data-driven clustering of the inter-model distances (Appendix D, Fig. C1). We average distances across seasons and member pairs to obtain a model-level distance matrix, and apply hierarchical agglomerative clustering with average linkage (Murtagh and Contreras, 2012), in the spirit of the model-genealogy analyses of Masson and Knutti (2011); Knutti et al. (2013). For the RMSE distance, the recovered hierarchy is largely consistent with the prior grouping: the MetUM, CESM2, CNRM and EC-Earth3 families emerge as compact clusters, indicating that the genealogical relationships of Table 1 are visible in the SLP fields themselves. The main exception is the MIROC family, whose two models are not grouped by either distance; here our prior partition is more conservative than the data-driven structure. Under the $1 - F_1$ distance, the family structure is resolved less sharply, and ACCESS-ESM1-5 separates from the rest of the MetUM family. However, this is consistent with the model employing an earlier-generation atmosphere than the other members of its family.



4.3 Validation of the reference-centrality assumption

Before running the pipeline for ClimWIP, we verify on the real ensemble the assumption made in the idealized model of Section 3, i.e. that the reference is not an outlier in the predictor space. We measure the position of ERA5 in the same condition
 520 in which each member enters a leave-one-family-out iteration of the perfect-model test, where a member is compared with all the members except those of its own family. For each of the 95 members coupled to an available projection we compute the mean distance to the members of the other families,

$$\bar{d}_i = \frac{1}{|\mathcal{T}_i|} \sum_{j \in \mathcal{T}_i} d(\mathbf{x}_i, \mathbf{x}_j), \quad \mathcal{T}_i = \{j : \mu(j) \neq \mu(i)\}, \quad (12)$$

using the family partition of Table 1. The same quantity is computed for the reference, $\bar{d}_{\text{ref}}$, whose average runs over all the
 525 members, since ERA5 belongs to no model family. We then report the rank of the reference in the ordering of the 96 objects (the 95 members plus ERA5) by increasing $\bar{d}$, together with the corresponding centrality percentile, namely the fraction of members for which $\bar{d}_i < \bar{d}_{\text{ref}}$. A reference lying in the bulk of the ensemble ranks about halfway through the ordering or earlier, one sitting at its very center ranks among the first, whereas an outlier, far from every model, would sit at its end. ERA5 ranks first for the RMSE, second for the SHD (1st percentile), fifth for the F_1 -distance (4th percentile) and 34th for the SID
 530 (35th percentile). The assumption that $\mathbf{x}_{\text{ref}}$ is not an outlier in predictor space is thus supported for all the distances we test.

4.4 Selection of the independence parameter σ_S

Using the family partition of Table 1, we select the independence parameter σ_S as the optimal threshold separating within-family and cross-family distances, following the criterion of Section 2.2. The results for each distance are reported in Fig. 4. We find that the RMSE and the F_1 -distance (panels a and b) provide the best separation between the two distributions. A
 535 good separation between the within-family and cross-family distance distributions indicates that the metric's geometry is consistent with the genealogy-based partition of Table 1. The immediate consequence concerns the identifiability of σ_S : when the two distributions are well separated, a threshold lying above typical within-family distances and below typical cross-family distances is well defined. However, we stress that separation is a property of the metric's similarity structure alone: it quantifies how well the metric recovers the assumed family partition and, thereby, whether the independence component of the scheme is
 540 well-posed for that metric. By itself, this does not imply that a well-separating metric yields more reliable weighted projections, since this depends on the relationship between projections and historical embeddings.

4.5 Choice of σ_D and weighting results

Next, we select the optimal performance parameter for each distance by applying coverage-based and CRPS-based perfect-model tests. Since we consider two separate projection periods (2041–2060 and 2081–2100) under the SSP5-8.5 scenario, an
 545 optimal σ_D is identified independently for each period. The quantities monitored in the two alternative tests are shown as a function of σ_D in Fig. 5.

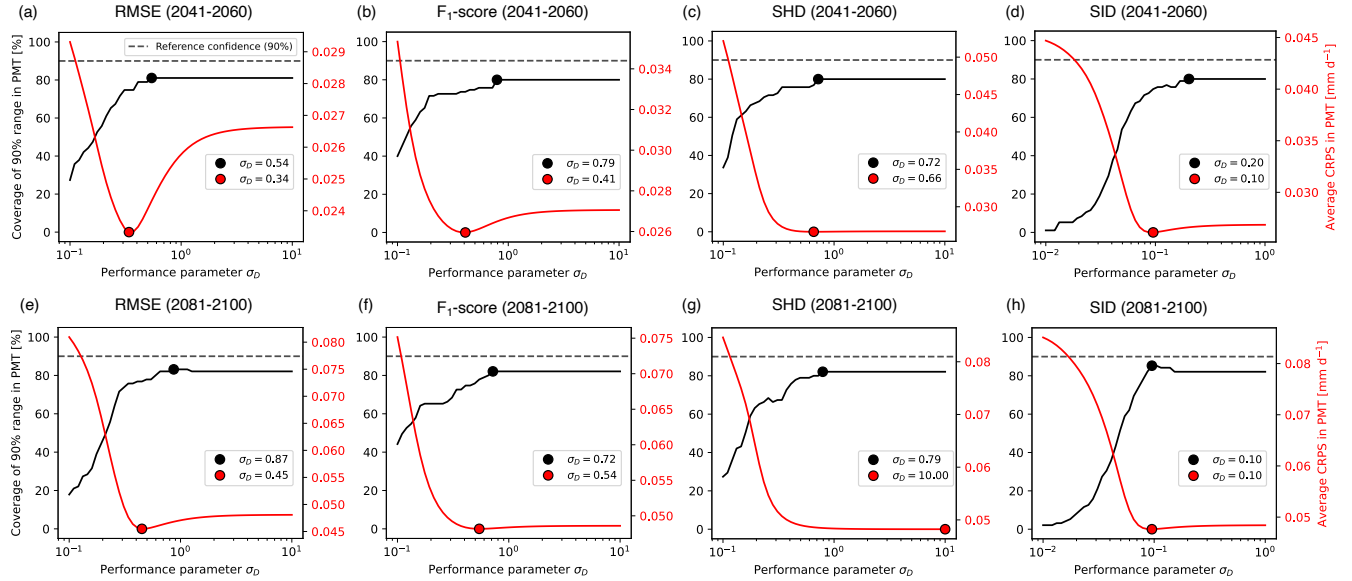


Figure 5. Selection of the performance parameter σ_D for the CMIP6 ensemble weighting. Each panel shows the two quantities monitored in the perfect-model test as a function of σ_D : the coverage of the 90% range (black, left axis) and the average CRPS between the weighted projection distribution and the pseudo-reference (red, right axis). The columns correspond to the four tested distances, and the rows to the two projection periods under SSP5-8.5: (a)–(d) 2041–2060 and (e)–(h) 2081–2100. The dashed line marks the nominal 90% confidence, which the coverage never attains. The black and red dots indicate the values of σ_D selected by the coverage-based and by the CRPS-based scheme, respectively, as reported in each panel. The abscissa is logarithmic; note the different range used for SID.

The coverage curves (black) display the same qualitative behavior for all distances and both periods: they rise steeply from zero, and then settle on a slightly lower plateau between 80% and 85%. The location of the maximum depends on the scale of the underlying distance: it falls at $\sigma_D \approx 0.5$ – 0.9 for the RMSE, the F_1 -distance and the SHD, but at $\sigma_D \approx 0.1$ – 0.2 for the SID, whose values live on a different scale. Crucially, the coverage always lies *below* the nominal 90% level (dashed lines), so that no value of σ_D satisfies the requirement of a non-overconfident coverage. This is the failure discussed in Section 2.3, which here appears for every metric and period considered: the prescription “the smallest σ_D for which the PMT coverage is at least nominal” admits no solution. To retain a coverage-based reference for comparison, we relax the criterion and select the σ_D that maximizes the PMT coverage, choosing the smallest σ_D in case of ties.

The CRPS curves (red) behave differently. For the RMSE (Fig. 5a,e) they exhibit a pronounced interior minimum, at $\sigma_D = 0.34$ and $\sigma_D = 0.45$ for the two periods, well separated from the $\sigma_D \rightarrow +\infty$ plateau. The depth of these minima is quantified by the PMT skill score of Eq. (9), reported in Fig. 6a. For both projection periods, the RMSE achieves by far the largest reduction in the PMT objective ($\approx -12\%$ and $\approx -5\%$ for the mid-century and end-of-century periods, respectively), indicating that the corresponding σ_D values are the most reliably identified. Interior minima are also found for the F_1 -score and SID, although progressively shallower. At the same time, for SHD the objective is almost flat: in the 2041–2060 period the minimum is

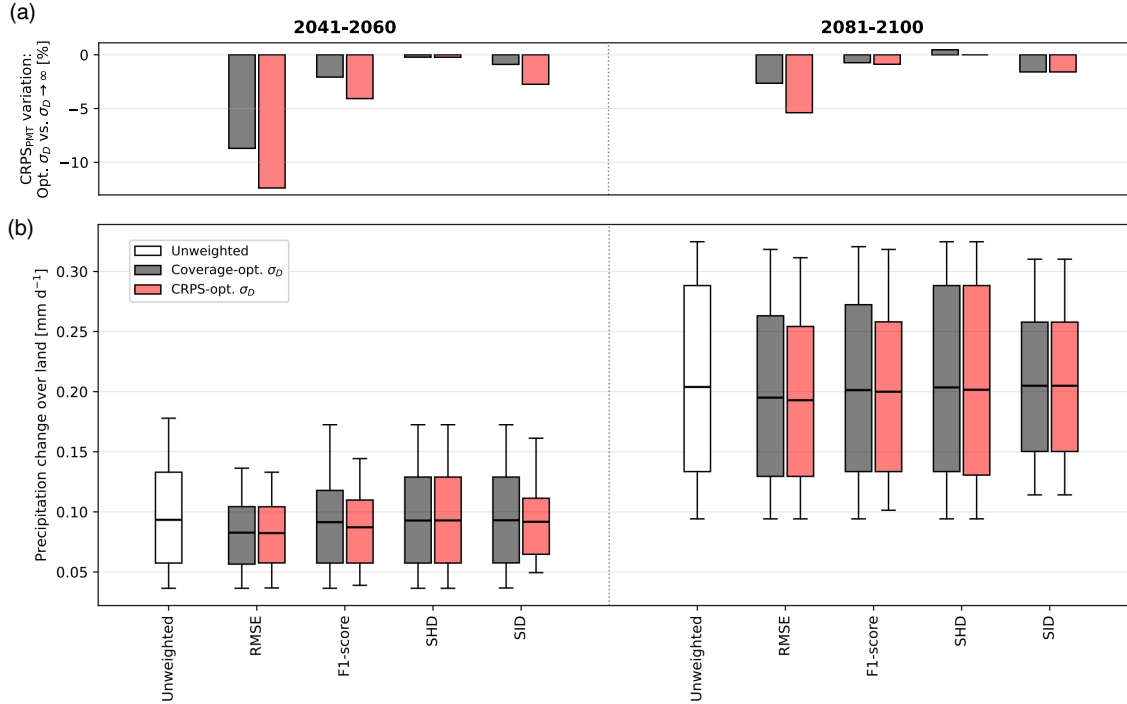


Figure 6. Weighting results for the CMIP6 ensemble. **(a)** PMT skill score of Eq. (9): relative variation of the average CRPS in the perfect-model test between the selected σ_D and the independence-only baseline ($\sigma_D \rightarrow +\infty$). Negative values indicate an improvement of the PMT objective; the more negative the value, the deeper the minimum from which σ_D is selected. **(b)** Projected change of precipitation over land with respect to the 1995–2014 baseline, for the unweighted ensemble and for the ensembles weighted with each of the tested distances. The boxes span the likely range (17th–83rd percentile), the whiskers the very likely range (5th–95th percentile), and the horizontal line marks the weighted mean. In both panels, grey and red denote the coverage-based and CRPS-based selection of σ_D , and the left and right blocks denote the 2041–2060 and 2081–2100 projection periods under SSP5-8.5.

barely distinguishable from the plateau, and in the 2081–2100 period it is not observed at all, as the CRPS objective decreases monotonically up to the upper boundary $\sigma_D = 10$. The corresponding skill scores in Fig. 6a are essentially zero, signaling that SHD carries little or no exploitable information for constraining the projections in both periods. We also note that the CRPS-optimal σ_D is systematically smaller than the coverage-optimal one, i.e. the CRPS-based scheme selects a more aggressive performance weighting. The two coincide only for SID in the 2081–2100 period (Fig. 5h), for which the two schemes return identical weights. For almost all distances, the PMT skill score of Fig. 6a is closer to zero in the 2081–2100 period than in 2041–2060, indicating that the performance parameter obtained for end-of-century projections is less sharply identified. We stress that this quantifies the signal the test is able to exploit, and not necessarily the strength of the resulting constraint: the latter is governed by the position of the minimiser $\sigma_{D, \text{opt.}}$, rather than by the depth of the minimum.



Table 2. Change of the projection distribution with respect to the unweighted ensemble. For each σ_D -selection scheme we report the shift of the ensemble mean, both as a percentage variation and in mm d^{-1} , and the relative change of the 66% and 90% interval widths (%). Negative values indicate a narrower (more constrained) interval and a downward shift of the mean. The values reported in bold text denote the largest constraints achieved according to each σ_D -selection scheme, for each confidence range and projection period.

Metric	Period	Coverage-opt. σ_D			CRPS-opt. σ_D		
		mean (%; mm d^{-1})	66% range	90% range	mean (%; mm d^{-1})	66% range	90% range
RMSE	2041–2060	–11.50% (–0.0107)	– 36.7%	– 29.4%	–11.24% (–0.0105)	– 43.0%	– 31.9%
	2081–2100	–4.36% (–0.0089)	–13.7%	–2.8%	–5.26% (–0.0107)	–19.4%	–5.8%
F ₁ -score	2041–2060	–2.08% (–0.0019)	–20.0%	–3.8%	–6.68% (–0.0062)	–30.6%	–25.5%
	2081–2100	–1.27% (–0.0026)	–10.3%	–1.8%	–1.93% (–0.0039)	–19.5%	–5.9%
SHD	2041–2060	–0.62% (–0.0006)	–5.4%	–3.8%	–0.55% (–0.0005)	–5.4%	–3.8%
	2081–2100	–0.17% (–0.0004)	0.0%	0.0%	–1.13% (–0.0023)	1.9%	0.0%
SID	2041–2060	–0.38% (–0.0004)	–5.5%	–4.0%	–1.75% (–0.0016)	–38.3%	–20.9%
	2081–2100	+0.51% (+0.0010)	– 30.5%	– 15.0%	+0.51% (+0.0010)	– 30.5%	– 15.0%

570 The weighted projections of precipitation change over land are shown in Fig. 6b, and the corresponding relative changes of the likely (17th–83rd percentile) and very likely (5th–95th percentile) ranges are reported in Table 2.

For the 2041–2060 period, the strongest constraints are obtained with the RMSE, which narrows the very likely range by 29.4% and 31.9% under the coverage- and CRPS-based selection, respectively. For the 2081–2100 period, the largest constraints are instead achieved by SID, which reduces the likely range by 30.5% and the very likely range by 15.0% under
 575 both schemes, against the 19.4% and 5.8% obtained with the RMSE. Consistently with its flat CRPS objective in the perfect-model tests, SHD produces no appreciable constraint in either period. Across all distances, the CRPS-based selection constrains more than the coverage-based one, as a direct consequence of the smaller σ_D it selects; the effect is most striking for SID in the 2041–2060 period, where the two schemes yield –38.3% and –5.5% on the likely range. Table 2 also reports the relative and absolute shifts of the means, which display a moderate variation according to all metrics, with the largest reduction achieved
 580 by the RMSE in the period 2041–2060 ($\simeq -11\%$, i.e. -0.011 mm d^{-1} , under both σ_D -selection schemes). Even the largest of these displacements amounts to less than a tenth of the corresponding very likely range, so the effect of the weighting is on the width of the ranges rather than on their center.

Read in isolation, these numbers would suggest that SID is the most effective distance for constraining end-of-century projections. The PMT skill score, however, warns against this conclusion, since the constraint achieved with SID rests on a
 585 far shallower CRPS minimum than that achieved with the RMSE. The comparison is clearest in the 2041–2060 period, where the two distances yield comparable reductions of the likely range (–43.0% for SID and –38.2% for the RMSE), while the underlying reduction of the PMT objective is about four times larger for the RMSE. Therefore, the PMT skill score indicates how justified a given constraint is, while the width of the weighted distribution gives the constraint itself. We finally remark that both observations corroborate the conclusion of the synthetic experiments (Fig. 2): projections further into the future are



590 harder to constrain, since the weakening of the time correlation with lead time reduces both the achievable constraint (with the only exception of SID using coverage-based weights, Table 2) and its reliability.

5 Conclusions

We have proposed a revision of the perfect-model test used to calibrate the performance parameter of climate projection weighting schemes such as ClimWIP, replacing the coverage of a pre-chosen confidence interval with the CRPS between the weighted projection distribution and the pseudo-reference. The change is conceptual before being technical. A coverage-based calibration asks only whether the pseudo-reality falls inside a stated interval. It is then combined with a rule (choosing the smallest admissible σ_D) whose explicit aim is to narrow that interval as much as the calibration tolerates: constraining is the objective, and predictive quality only a boundary condition. Scoring the full predictive distribution through the CRPS reverses this order. The CRPS is optimized in expectation by a distribution that is well centered on the target and neither too dispersed nor too narrow. The resulting weighting is the one expected to predict best, and the narrowing of the projection range follows as a consequence rather than being pursued for its own sake.

This shift also removes a practical failure of the coverage criterion, which we first identified on synthetic data and then observed throughout our CMIP6 application. The prescription “the smallest σ_D whose coverage is at least nominal” presupposes that such a σ_D exists, and this is not guaranteed. In several cases we examined, the PMT coverage saturated below the nominal level for every σ_D , leaving the criterion without a solution. In contrast, the minimum of CRPS always remains a well-posed target. This minimum need not be interior: for example, for SHD at the end of the century it lies at the upper boundary, informing that the predictor distance carries no signal that the performance weighting can exploit.

Our results show that the CRPS-based scheme constrains the projections more than its coverage-based counterpart, and this comparison deserves a careful reading. Part of the difference stems from our own conservative handling of the infeasible coverage constraint. In the absence of an admissible σ_D , we retained the smallest value of σ_D maximizing the coverage, which lies on the saturated portion of the curve and therefore corresponds to a weak weighting. Comparable, or even stronger, constraints could be obtained from the coverage criterion by deliberately accepting an empirical coverage below the nominal level (e.g. Brunner et al., 2019; Debeire et al., 2025). The difficulty is that such a relaxation is arbitrary: no quantitative criterion establishes how far below the nominal level one may go, and the amount of tolerated overconfidence becomes a free choice of the analyst. The main advantage of the CRPS-based scheme is therefore not that it constrains more, but that it allows constraining through a well-defined optimization criterion.

Two further contributions follow from the same construction. First, the skill score (i.e. the relative reduction of the CRPS achieved at the optimum) provides a quality measure of the calibration itself: a pronounced reduction indicates that the perfect-model test locates a sharply defined optimum, whereas a value close to zero reveals an objective that is nearly flat in σ_D , whose minimizer, as well as the resulting weighting, is poorly determined. Since such a relative reduction is computed as a by-product of the calibration, it can be used to compare the reliability of alternative predictor distances before committing to



any. Second, our synthetic experiments show that the CRPS-based scheme is more sensitive than the coverage-based one to a misspecification of the model families. A conservative definition of the families is therefore necessary for its use.

Applied to CMIP6 projections of precipitation over land under SSP5-8.5, the framework yields a substantial constraint for the 2041–2060 period, and a considerably weaker one at the end of the century, where the perfect-model test extracts little signal from the historical SLP fields for any of the distances considered. Among these, the distances built on causal-network representations do carry information (for example, SID achieves the largest reduction of the projection ranges for the 2081–2100 period) but the skill score tempers this result: the same constraint is supported by a far shallower optimum than the one underlying the area-weighted RMSE, which remains the most reliable predictor distance in both periods despite being by far the simplest. That a large constraint and a well-supported constraint need not coincide is, in our view, the main practical lesson of this application.

Several directions remain open. The most immediate one follows from the smoothness of the objective: since the CRPS is defined and continuous for any pair (σ_S, σ_D) , it could be minimized jointly over both parameters, replacing the sequential calibration adopted here, in which σ_S is fixed first through a heuristic separation of within- and cross-family distances, by a single optimization of a scoring rule. Joint calibration over (σ_S, σ_D) has precedent (Lorenz et al., 2018), though with coverage as the criterion. More generally, nothing in the CRPS-based construction of perfect-model tests is specific to the weighting function of Eq. (1): the same objective could calibrate the free parameters of other weighting schemes, and the same skill score could assess their reliability. Extending the approach to spatially resolved projections would instead require replacing the CRPS with a multivariate scoring rule, and combining several predictors into a multi-diagnostic weighting remains a natural development. Finally, the three graph distances we test are computed on the same causal graphs, yet they yield markedly different constraints and skill scores. This indicates that the way graphs are compared matters at least as much as the causal representation itself. Distances that weight links by their estimated strength (Runge et al., 2019b), for which metrics on weighted directed networks are available (Chowdhury and Mémoli, 2019), or that restrict the interventional comparison underlying the SID to relevant subsets of nodes, are a natural avenue.

Code and data availability. The CMIP6 historical and SSP5-8.5 simulations used here are publicly available through the Earth System Grid Federation (Cinquini et al., 2014); the complete list of models and members, together with the stage of the analysis each one enters, is given in Table S1 in the Supplement. The ERA5 reanalysis is distributed by the Copernicus Climate Change Service (Hersbach et al., 2020). The causal graphs and the preprocessing pipeline that produced them were released by Debeire (2025, code). The data supporting the present work (the precipitation projections and causal graphs of Debeire (2025), reused here, and the monthly sea-level pressure fields produced for this study together with the ESMValTool recipe that generated them) are archived on Zenodo at <https://doi.org/10.5281/zenodo.22775441>. The code implementing the coverage-based and CRPS-based perfect-model tests, the skill score of Eq. (9), and the synthetic-data experiments of Sect. 3 is available at <https://github.com/vdeltatto/climwip-crps>.



Appendix A: Algorithmic pipeline for perfect-model tests

In Algorithm A1 we report the coverage-based implementation of the perfect-model test for a single candidate value of the performance parameter σ_D . We denote by $\mu(i)$ the family to which member i belongs, so that two members i and j are considered redundant when $\mu(i) = \mu(j)$; the leave-one-family-out step removes, for each pseudo-reference k , all members sharing its family (i.e. all i with $\mu(i) = \mu(k)$). The nominal level is denoted by q (e.g. $q = 0.9$ for a 5th–95th percentile interval), and the output is the PMT coverage $c(\sigma_D)$, the fraction of pseudo-references whose projection falls within the weighted confidence interval. The performance parameter is then selected by repeating the procedure over a range of σ_D values, as described in Section 2.3. We write N for the number of members with an available projection, so that the loop runs over all N pseudo-references, and M for the number of survivors in iteration k ; the latter varies with k because families differ in size, but we leave this dependence implicit.

Algorithm A1 Coverage-based perfect-model test for a candidate σ_D

```

1: Input: representations  $\{\mathbf{x}_i\}_{i=1}^M$  and projections  $\{p_i\}_{i=1}^M$ ; family label  $\mu(i)$  for each member; frozen  $\sigma_S$ ; candidate  $\sigma_D$ ; nominal level  $q$ .
2: Output: PMT coverage  $c(\sigma_D)$ .
3:  $n_{\text{in}} \leftarrow 0$ 
4: for  $k = 1$  to  $M$  do
5:    $\mathcal{T} \leftarrow \{i : \mu(i) \neq \mu(k)\}$  {leave-one-family-out training set}
6:   for  $i \in \mathcal{T}$  do
7:      $w_i \leftarrow \frac{e^{-d(\mathbf{x}_i, \mathbf{x}_k)^2 / \sigma_D^2}}{1 + \sum_{j \in \mathcal{T}, j \neq i} e^{-d(\mathbf{x}_i, \mathbf{x}_j)^2 / \sigma_S^2}}$ 
8:   end for
9:   normalize  $\{w_i\}_{i \in \mathcal{T}}$  such that  $\sum_{i \in \mathcal{T}} w_i = 1$ 
10:   $[\ell_k, u_k] \leftarrow$  weighted  $q$ -confidence interval of  $\{p_i\}_{i \in \mathcal{T}}$ 
11:  if  $\ell_k \leq p_k \leq u_k$  then
12:     $n_{\text{in}} \leftarrow n_{\text{in}} + 1$ 
13:  end if
14: end for
15:  $c(\sigma_D) \leftarrow n_{\text{in}} / M$ 
16: return  $c(\sigma_D)$ 

```

The interval computed at line 10 of Algorithm A1 is defined by a pair of weighted quantiles (ℓ_k and u_k), which we define explicitly in the following following the implementation in `statsmodels` (Seabold and Perktold, 2010). Let the M surviving projections of iteration k be labelled in increasing order, $p_1 \leq \dots \leq p_M$, with w_l the weight attached to p_l and $\sum_l w_l = 1$ (projections that coincide exactly are merged beforehand, summing their weights), and let $W_l = \sum_{l' \leq l} w_{l'}$ be the cumulative weight. For a level $\theta \in (0, 1)$, we define the weighted θ -quantile $\hat{Q}(\theta)$ as the smallest projection whose cumulative weight reaches θ , namely $\hat{Q}(\theta) = p_{l^*}$ with $l^* = \min\{l : W_l \geq \theta\}$. In the boundary case $W_{l^*} = \theta$ with $l^* < M$, the two adja-



cent projections are averaged, $\hat{Q}(\theta) = \frac{1}{2}(p_{l^*} + p_{l^*+1})$. The endpoints of the q -confidence interval are then $\ell_k = \hat{Q}(\frac{1-q}{2})$ and
 670 $u_k = \hat{Q}(\frac{1+q}{2})$, so that $q = 0.9$ returns the 5th and 95th weighted percentiles.

Algorithm A2 reports the CRPS-based implementation. It differs from the coverage-based test (Algorithm A1) only in the per-iteration evaluation: rather than recording whether the pseudo-reference falls inside a confidence interval, it computes the CRPS between the weighted projection distribution and the pseudo-reference. The performance parameter is then selected as the minimizer $\sigma_D^* = \arg \min_{\sigma_D} \langle \text{CRPS}(\sigma_D) \rangle$, rather than as the smallest σ_D satisfying a coverage constraint.

Algorithm A2 CRPS-based perfect-model test for a candidate σ_D

```

1: Input: as in Algorithm A1, except the nominal level  $q$  is not required.
2: Output: average CRPS,  $\langle \text{CRPS}(\sigma_D) \rangle$ 
3:  $S \leftarrow 0$ 
4: for  $k = 1$  to  $M$  do
5:    $\mathcal{T} \leftarrow \{i : \mu(i) \neq \mu(k)\}$  {leave-one-family-out training set}
6:   for  $i \in \mathcal{T}$  do
7:      $w_i \leftarrow \frac{e^{-d(\mathbf{x}_i, \mathbf{x}_k)^2 / \sigma_D^2}}{1 + \sum_{j \in \mathcal{T}, j \neq i} e^{-d(\mathbf{x}_i, \mathbf{x}_j)^2 / \sigma_D^2}}$ 
8:   end for
9:   normalize  $\{w_i\}_{i \in \mathcal{T}}$  so that  $\sum_{i \in \mathcal{T}} w_i = 1$ 
10:   $\text{CRPS}_k \leftarrow \sum_{i \in \mathcal{T}} w_i |p_i - p_k| - \frac{1}{2} \sum_{i \in \mathcal{T}} \sum_{j \in \mathcal{T}} w_i w_j |p_i - p_j|$  {CRPS between weighted distribution and pseudo-reference  $p_k$ }
11:   $S \leftarrow S + \text{CRPS}_k$ 
12: end for
13:  $\langle \text{CRPS}(\sigma_D) \rangle \leftarrow S/M$ 
14: return  $\langle \text{CRPS}(\sigma_D) \rangle$ 
    
```

675 Appendix B: Details on the synthetic data model

The representation space used for the experiments of Section 3 is a d -dimensional Euclidean space (we use $d = 3$), and each ensemble member is a point $\mathbf{x}_i \in \mathbb{R}^d$. The members are organized into K families with a variable number of members: $K = 4$ with 25 members per family in the example of Fig. 1 and in the tests of Figs. 2a-c and Fig. 3, $K = 12$ with the uneven split of Table 1 in the tests of Figs. 2d-f. Each family μ has an “initial” centroid $\mathbf{c}_\mu(t_1)$; the centroids are drawn independently and
 680 uniformly in the box $[0, L]^d$ with $L = 10$. All families share the same isotropic covariance $\xi^2 I_d$, with $\xi = 1.5$. The common isotropic spread, together with the scale of the centroid separations, is chosen so that the resulting within-family and cross-family distance distributions overlap to a degree comparable with the CMIP6 ensemble (Fig. 4).

We interpret the representation $\mathbf{x}_i$ as the historical state $\mathbf{h}_i(t_1)$ of member i , sampled as

$$\mathbf{x}_i = \mathbf{h}_i(t_1) = \mathbf{c}_{\mu(i)}(t_1) + \boldsymbol{\eta}_i(t_1), \quad \boldsymbol{\eta}_i(t_1) \sim \mathcal{N}(\mathbf{0}, \xi^2 I_d). \quad (\text{B1})$$



685 To generate the future states $\mathbf{h}_i(t_2)$, we first compute the future centroids,

$$\mathbf{c}_\mu(t_2) = \rho \mathbf{c}_\mu(t_1) + \sqrt{1 - \rho^2} \boldsymbol{\zeta}_\mu, \quad (\text{B2})$$

where $\{\boldsymbol{\zeta}_\mu\}$ is a set of independent random displacements drawn from the same distribution as $\{\mathbf{c}_\mu(t_1)\}$, representing the motion of the centroid position over the projection horizon. The member-specific fluctuations evolve according to the same form,

690 $\boldsymbol{\eta}_i(t_2) = \rho \boldsymbol{\eta}_i(t_1) + \sqrt{1 - \rho^2} \boldsymbol{\eta}'_i, \quad (\text{B3})$

where $\boldsymbol{\eta}'_i \sim \mathcal{N}(\mathbf{0}, \xi^2 I_d)$ independently of $\boldsymbol{\eta}_i(t_1)$. The future state is then simply

$$\mathbf{h}_i(t_2) = \mathbf{c}_{\mu(i)}(t_2) + \boldsymbol{\eta}_i(t_2). \quad (\text{B4})$$

The parameter $\rho \in [0, 1]$ is the autocorrelation shared by the two components of each member's state: the family centroid and the member-specific fluctuation around it. The factors $\sqrt{1 - \rho^2}$ keep the marginal variance unchanged between t_1 and
695 t_2 , so that the within-family spread and the between-family centroid spread are held constant regardless of ρ . This parameter can be implicitly interpreted as a function of the projection time horizon: the autocorrelation is maximum ($\rho = 1$) in the limit of identical historical and projection times ($t_1 = t_2$), whereas it is minimum ($\rho \rightarrow 0$) in the limit of a long projection horizon ($|t_2 - t_1| \rightarrow +\infty$). In the former case the centroids are frozen at $\mathbf{c}_\mu(t_2) = \mathbf{c}_\mu(t_1)$, and $\mathbf{h}_i(t_2) = \mathbf{h}_i(t_1)$; in the latter, both member-specific fluctuations and family centroid positions at t_2 become fully independent of their values at t_1 . Idealized
700 settings of this kind, based on linear stochastic models with prescribed time scales, have been used to study detectability and predictability in ensembles of climate simulations (Kravtsov et al., 2022; Gavrilov et al., 2024).

The scalar projection is simply taken as the first coordinate of the future state,

$$p_i = \mathbf{h}_i(t_2) \cdot \hat{\mathbf{e}}_1, \quad \hat{\mathbf{e}}_1 = (1, 0, 0). \quad (\text{B5})$$

By the isotropy of the dynamics, any other fixed direction would be statistically equivalent. The reference pair $(\mathbf{x}_{\text{ref}}, p_{\text{ref}})$
705 is generated according to the same equations, but with $\mathbf{c}_{\text{ref}}(t_1) = \frac{1}{K} \sum_\mu \mathbf{c}_\mu$ and $\boldsymbol{\zeta}_{\text{ref}} = \frac{1}{K} \sum_\mu \boldsymbol{\zeta}_\mu$. The first choice places the reference in the bulk of the ensemble at the historical time t_1 ; the second guarantees that it remains there at the projection time t_2 .

For the systematic experiments (Figs. 2 and 3), the centroids $\{\mathbf{c}_\mu\}$, the centroid displacements $\{\boldsymbol{\zeta}_\mu\}$, the member fluctuations, and the reference are re-drawn independently. The results are averaged over 200 realizations for each tested value of
710 ρ .

Appendix C: Graph-based distances

The graphs are inferred with the PCMCI causal discovery algorithm (Runge et al., 2019b), which takes as input a multivariate time series and returns a time-series causal graph encoding the lagged dependencies among its components. In the pipeline of



Debeire et al. (2025), these components are the modes obtained from the dimensionality reduction of the SLP field, of which 60
715 are retained, and the maximum lag is set to $\tau_{\max} = 20$ d. Since we work with the same graph dataset, our representations $\mathbf{x}_i^s$ are
time-series causal graphs over 60 nodes and time lags $\tau \in \{0, 1, \dots, 20\}$, encoding the causal links for each season s separately.
We take the first slice as the present, $\tau = 0$, and the remaining slices as future times $\tau \geq 1$. A directed link $(a, \tau_1) \rightarrow (b, \tau_2)$
with $\tau_1 < \tau_2$ encodes a causal dependence from mode a to mode b with lag $\tau_2 - \tau_1$; since links point forward in time, the
graph is a directed acyclic graph (DAG). By stationarity, the graph is fully specified by the links originating in the present slice,
720 $(a, 0) \rightarrow (b, \tau)$ with $\tau \geq 1$, which we call the *present-to-future* links: this is the natural set on which to compare the presence
or absence of individual causal links. The full time-series graph, however, is obtained by repeating these links periodically
over consecutive slices (in principle infinitely) so that it explicitly represents links between any two lagged slices (a, τ_1) and
 (b, τ_2) with $1 \leq \tau_1 < \tau_2$. These repeated links make the composite paths $(a, 0) \rightarrow (c, \tau_1) \rightarrow (b, \tau_2)$ explicit. Such indirect paths
are essential when the graphs are compared through causal metrics: the causal effect of one node on another is mediated by
725 all the directed paths connecting them, whereas in a representation restricted to the present-to-future links no node at $\tau \geq 1$
has outgoing edges, and no path longer than a single link can exist. Two graphs can therefore be compared either over their
present-to-future links alone, or over the full “unrolled” representation, which we truncate at the $\tau_{\max} + 1 = 21$ slices available.
The two options differ substantially in size: in the first case the comparison involves the $60 \times 60 \times 20$ binary tensor of possible
present-to-future links, whereas in the second the graph is described by the adjacency matrix of a DAG with $60 \times 21 = 1260$
730 nodes (one per mode and per slice), i.e. a 1260×1260 binary matrix. We denote the graphs being compared by $\mathcal{G}_{\text{ref}}$ and $\mathcal{G}$, and
their binary adjacency (over the relevant link set) by A_{ref} and A . All distances are computed per season and averaged over the
four seasons. Moreover, they are symmetrized before being used within the ClimWIP scheme (Eq. (1)).

C1 F_1 -score distance

The F_1 -score compares the two graphs over the set of present-to-future links $(a, 0) \rightarrow (b, \tau)$, cast as a binary classification
735 in which the links of $\mathcal{G}$ are the “predictions” and those of $\mathcal{G}_{\text{ref}}$ the “ground truth”. With TP the links present in both graphs,
FP those present only in $\mathcal{G}$, and FN those present only in $\mathcal{G}_{\text{ref}}$, precision and recall are $\text{TP}/(\text{TP} + \text{FP})$ and $\text{TP}/(\text{TP} + \text{FN})$,
respectively, and

$$F_1(\mathcal{G}_{\text{ref}}, \mathcal{G}) = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \quad (\text{C1})$$

The score is defined between 0 and 1: it equals 1 when the two link sets coincide ($\text{FP} = \text{FN} = 0$), and 0 when the two graphs
740 share no link ($\text{TP} = 0$). We convert it into a distance as

$$d_{F_1}(\mathcal{G}_{\text{ref}}, \mathcal{G}) = 1 - F_1(\mathcal{G}_{\text{ref}}, \mathcal{G}), \quad (\text{C2})$$

which vanishes for identical link sets.

Following the implementation in Debeire (2025, code), we do not adopt the raw definition above, but introduce three modifi-
cations. First, a tolerance $\tau_{\text{tol}} = 2$ d is allowed on the lag when matching two links: a link $(a, 0) \rightarrow (b, \tau)$ of $\mathcal{G}_{\text{ref}}$ is counted as a
745 true positive if $\mathcal{G}$ contains a link $(a, 0) \rightarrow (b, \tau')$ with $|\tau' - \tau| \leq \tau_{\text{tol}}$, which mitigates the sensitivity to the precise inferred lags.

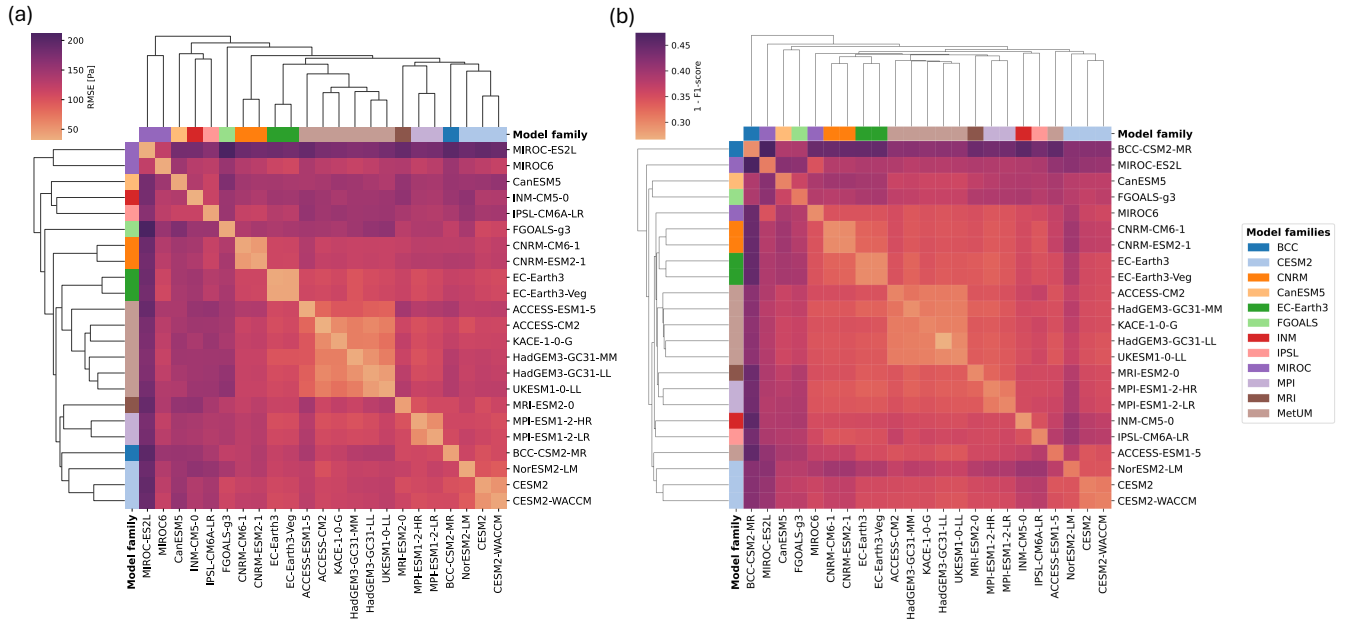


Figure C1. Data-driven validation of the model-family partition. Panels (a) and (b) show the model–model distance matrices for the area-weighted RMSE and the F_1 -score distance, respectively, with the dendrograms obtained from an average-linkage hierarchical clustering of the same matrices. Each entry is the corresponding distance averaged over the four seasons and over all pairs of members of the two models; diagonal entries report the average intra-model distance between distinct members of the same model. Rows and columns are ordered according to the clustering, and the side bars encode the prior family assigned to each model in Table 1, so that data-driven clusters and prior families can be compared directly.

Second, a match additionally requires the two links to carry a dependence of the same sign, i.e. the partial correlations associated with the link in $\mathcal{G}$ and in $\mathcal{G}_{\text{ref}}$ must be either both positive or both negative, so that links encoding opposite dependencies are not counted as agreeing. Third, self-links $(a, 0) \rightarrow (a, \tau)$ are excluded from the counting.

C2 Structural Hamming distance (SHD)

750 The Structural Hamming distance (Tsamardinos et al., 2006) counts the links on which the two graphs disagree, over the same set of lagged links into the present:

$$d_{\text{SHD}}(\mathcal{G}_{\text{ref}}, \mathcal{G}) = \sum_{a,b,\tau} |A_{\text{ref}}^{ab}(\tau) - A^{ab}(\tau)|, \quad (\text{C3})$$

where $A^{ab}(\tau)$ is 1 when the link $(a, 0) \rightarrow (b, \tau)$ is present, and 0 otherwise. It is the total number of link insertions and deletions needed to turn one graph into the other. The usual third contribution to the SHD, the count of links whose orientation differs
755 between the two graphs, is vacuous here: all present-to-future links point forward in time, so the reverse of a link is never itself an admissible link and no orientation mismatch can occur. Equation (C3) is therefore the full SHD for this class of graphs and



not a simplification of it. Like the F_1 -score, the SHD is a purely structural comparison: it inspects the presence or absence of individual links. It is therefore well defined on the lagged-into-present link set alone, without reference to the paths connecting the links.

760 C3 Structural Intervention distance (SID)

The Structural Intervention distance (Peters and Bühlmann, 2015) compares the graphs through their causal content rather than their individual links. Estimating the causal effect of a on b from observational data requires conditioning on a suitable set of other nodes, called the *adjustment set*, chosen to block the non-causal paths between a and b . Given such a set, a DAG determines the interventional distribution of b under an intervention on a . Adjustment sets between pairs of nodes are not
 765 uniquely determined, and a valid criterion to build them is *parent adjustment*: for each ordered pair (a, b) , the adjustment set is taken to be the set of parents of a . SID counts the ordered pairs (a, b) for which the parents of a in $\mathcal{G}$ do not form a valid adjustment set in $\mathcal{G}_{\text{ref}}$, and would therefore yield a wrong interventional distribution. Therefore, it measures how often a graph used in place of the other would lead to an incorrect causal-effect estimate. Because the validity of an adjustment set is a property of the whole graph and not of individual links, SID must be evaluated on the full unrolled graph, including the links
 770 $(a, \tau_1) \rightarrow (b, \tau_2)$ between two lagged slices. For its practical computation, we employed the Python library `gadjid` (Henckel et al., 2024).

Appendix D: Validation of the family partition through hierarchical clustering

To complement the genealogy-based definition of the model families (Table 1) with a data-driven validation, we perform a hierarchical cluster analysis of the inter-model distances, following the approach used to characterize model genealogy in
 775 previous CMIP studies (Masson and Knutti, 2011; Knutti et al., 2013).

We first reduce the member-level distances to a model-level distance matrix: for each pair of models, the distance is averaged over the four seasons and over all pairs of members belonging to the two models, yielding a symmetric 23×23 matrix. The diagonal, which contains the average distance between distinct members of the same model, is shown in the matrix but is not used by the clustering algorithm, which operates only on inter-model distances. We then apply an agglomerative hierarchical
 780 clustering with average linkage (UPGMA, Murtagh and Contreras 2012), in which the distance between two groups is the mean distance over all pairs of models belonging to them. The resulting dendrogram is shown in Fig. C1 alongside the re-ordered distance matrix, with a color bar encoding the prior family of each model, so that the agreement between the data-driven structure and the partition of Table 1 can be inspected directly. The analysis is repeated for the area-weighted RMSE and for the $1 - F_1$ distance. The computations are performed with the `scipy` implementation of the linkage and dendrogram
 785 algorithms (Virtanen et al., 2020).



Author contributions. VDT designed the study, implemented the CRPS-based perfect-model test and the synthetic-data experiments, carried out the analysis, and produced the figures. AG contributed to the design of the synthetic data model and to the interpretation of the CMIP6 results. GCV supervised the work and secured funding. VDT wrote the first draft; all authors discussed the results and revised the manuscript.

Competing interests. The authors declare that they have no competing interests.

790 *Acknowledgements.* We thank Kevin Debeire and colleagues for sharing the causal-graph and the projection dataset employed in Debeire et al. (2025), which made the comparison of predictor distances in Sec. 4 possible. We acknowledge the World Climate Research Programme, which coordinated CMIP6, the climate modeling groups listed in Table 1 for producing and sharing their simulations, and the Earth System Grid Federation for archiving and distributing them. This work is supported by the European Research Council (ERC) through grant USMILE (grant agreement No. 855187); the European Commission through the Horizon Europe projects ELIAS (grant agreement No. 101120237)
795 and AI4PEX (grant agreement No. 101137682). While all intellectual work and writing has been performed by the authors, Claude Opus 4.8 and 5.0 (Anthropic) were used to improve the English of parts of the text.



References

- Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, *Earth System Dynamics*, 10, 91–105, <https://doi.org/10.5194/esd-10-91-2019>, 2019.
- Andela, B., Broetz, B., de Mora, L., Drost, N., Eyring, V., Koldunov, N., Lauer, A., Mueller, B., Predoi, V., Righi, M., Schlund, M., Vegas-Regidor, J., Zimmermann, K., Adeniyi, K., Castellani, G., Arnone, E., Bellprat, O., Berg, P., Billows, C., Blockley, E., Bock, L., Bodas-Salcedo, A., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Corti, S., Crezee, B., Davin, E. L., Davini, P., Deser, C., Diblen, F., Docquier, D., Dreyer, L., Ehbrecht, C., Earnshaw, P., Geddes, T., Gier, B., Gillett, E., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hall, S., Hardacre, C., von Hardenberg, J., Hassler, B., Heuer, H., Hogan, E., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Kuehbach, B., Lledó, L., Lejeune, Q., Lembo, V., Little, B., Loosveldt-Tomas, S., Lorenz, R., Lovato, T., Lucarini, V., Malinina, E., Massonnet, F., Mohr, C. W., Amarjiit, P., Parsons, N., Pérez-Zanón, N., Phillips, A., Proft, M., Russell, J., Sandstad, M., Sellar, A., Senftleben, D., Serva, F., Sillmann, J., Stacke, T., Storkey, D., Swaminathan, R., Tomkins, K., Torralba, V., Weigel, K., Sarauer, E., Schulze, K., Roberts, C., Kalverla, P., Alidoost, S., Verhoeven, S., Vreede, B., Smeets, S., Soares Siqueira, A., Kazeroni, R., Potter, J., Winterstein, F., Beucher, R., Kraft, J., Ruhe, L., Bonnet, P., Munday, G., Chun, F., and Ellis, H.: ESMValTool, <https://doi.org/10.5281/zenodo.18998400>, zenodo [code], 2026.
- Annan, J. D. and Hargreaves, J. C.: On the meaning of independence in climate science, *Earth System Dynamics*, 8, 211–224, <https://doi.org/10.5194/esd-8-211-2017>, 2017.
- Bi, D., Dix, M., Marsland, S., O’Farrell, S., Sullivan, A., Bodman, R., Law, R., Harman, I., Srbinovsky, J., Rashid, H. A., Dobrohotoff, P., Mackallah, C., Yan, H., Hirst, A., Savita, A., Dias, F. B., Woodhouse, M., Fiedler, R., and Heerdegen, A.: Configuration and spin-up of ACCESS-CM2, the new generation Australian Community Climate and Earth System Simulator Coupled Model, *Journal of Southern Hemisphere Earth Systems Science*, 70, 225–251, <https://doi.org/10.1071/ES19040>, 2020.
- Boé, J.: Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity, *Geophysical Research Letters*, 45, 2771–2779, <https://doi.org/10.1002/2017GL076829>, 2018.
- Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R.: Quantifying uncertainty in European climate projections using combined performance-independence weighting, *Environmental Research Letters*, 14, 124 010, <https://doi.org/10.1088/1748-9326/ab492f>, 2019.
- Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., and Knutti, R.: Reduced global warming from CMIP6 projections when weighting models by performance and independence, *Earth System Dynamics*, 11, 995–1012, <https://doi.org/10.5194/esd-11-995-2020>, 2020.
- Camps-Valls, G., Gerhardus, A., Ninad, U., Varando, G., Martius, G., Balaguer-Ballester, E., Vinuesa, R., Diaz, E., Zanna, L., and Runge, J.: Discovering causal relations and equations from data, *Physics Reports*, 1044, 1–68, <https://doi.org/10.1016/j.physrep.2023.10.005>, 2023.
- Chowdhury, S. and Mémoli, F.: The Gromov–Wasserstein distance between networks and stable network invariants, *Information and Inference: A Journal of the IMA*, 8, 757–787, <https://doi.org/10.1093/imaia/iaz026>, 2019.
- Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., Ananthakrishnan, R., Miller, N., Denvil, S., Morgan, M., Pobre, Z., Bell, G. M., Doutriaux, C., Drach, R., Williams, D., Kershaw, P., Pascoe, S., Gonzalez, E., Fiore, S., and Schweitzer, R.: The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data, *Future Generation Computer Systems*, 36, 400–417, <https://doi.org/10.1016/j.future.2013.07.002>, special Section: Intelligent Big Data Processing Special Section: Behavior Data



- Security Issues in Network Information Propagation Special Section: Energy-efficiency in Large Distributed Computing Architectures
Special Section: eScience Infrastructure and Applications, 2014.
- 835 Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R.,
Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb,
W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein,
M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C.,
Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2
840 (CESM2), *Journal of Advances in Modeling Earth Systems*, 12, e2019MS001916, <https://doi.org/10.1029/2019MS001916>, 2020.
- Debeire, K.: Constraining uncertainty in projected precipitation over land with causal discovery, <https://doi.org/10.5281/zenodo.14865765>,
zenodo [code], 2025.
- Debeire, K., Bock, L., Nowack, P., Runge, J., and Eyring, V.: Constraining uncertainty in projected precipitation over land with causal
discovery, *Earth System Dynamics*, 16, 607–630, <https://doi.org/10.5194/esd-16-607-2025>, 2025.
- 845 Deser, C., Knutti, R., Solomon, S., and Phillips, A. S.: Communication of the role of natural variability in future North American climate,
Nature Climate Change, 2, 775–779, <https://doi.org/10.1038/nclimate1562>, 2012a.
- Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: the role of internal variability, *Climate
Dynamics*, 38, 527–546, <https://doi.org/10.1007/s00382-010-0977-x>, 2012b.
- Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., Bernardello, R., Boussetta, S., Caron, L.-P., Carver, G.,
850 Castrillo, M., Catalano, F., Cvijanovic, I., Davini, P., Dekker, E., Doblas-Reyes, F. J., Docquier, D., Echevarria, P., Fladrich, U., Fuentes-
Franco, R., Gröger, M., v. Hardenberg, J., Hieronymus, J., Karami, M. P., Keskinen, J.-P., Koenigk, T., Makkonen, R., Massonnet, F.,
Ménégoz, M., Miller, P. A., Moreno-Chamarro, E., Nieradzick, L., van Noije, T., Nolan, P., O'Donnell, D., Ollinaho, P., van den Oord,
G., Ortega, P., Prims, O. T., Ramos, A., Reerink, T., Rousset, C., Ruprich-Robert, Y., Le Sager, P., Schmith, T., Schrödner, R., Serva, F.,
Sicardi, V., Sloth Madsen, M., Smith, B., Tian, T., Tourigny, E., Uotila, P., Vancoppenolle, M., Wang, S., Wårlind, D., Willén, U., Wyser,
855 K., Yang, S., Yepes-Arbós, X., and Zhang, Q.: The EC-Earth3 Earth system model for the Coupled Model Intercomparison Project 6,
Geoscientific Model Development, 15, 2973–3020, <https://doi.org/10.5194/gmd-15-2973-2022>, 2022.
- Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model
Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geoscientific Model Development*, 9, 1937–1958,
<https://doi.org/10.5194/gmd-9-1937-2016>, 2016.
- 860 Eyring, V., Collins, W. D., Gentile, P., Barnes, E. A., Barreiro, M., Beucler, T., Bocquet, M., Bretherton, C. S., Christensen, H. M.,
Dagon, K., Gagne, D. J., Hall, D., Hammerling, D., Hoyer, S., Iglesias-Suarez, F., Lopez-Gomez, I., McGraw, M. C., Meehl, G. A.,
Molina, M. J., Monteleoni, C., Mueller, J., Pritchard, M. S., Rolnick, D., Runge, J., Stier, P., Watt-Meyer, O., Weigel, K., Yu, R.,
and Zanna, L.: Pushing the frontiers in climate modelling and analysis with machine learning, *Nature Climate Change*, 14, 916–928,
<https://doi.org/10.1038/s41558-024-02095-y>, 2024a.
- 865 Eyring, V., Gentile, P., Camps-Valls, G., Lawrence, D. M., and Reichstein, M.: AI-empowered next-generation multiscale climate modelling
for mitigation and adaptation, *Nature Geoscience*, 17, 963–971, <https://doi.org/10.1038/s41561-024-01527-w>, 2024b.
- Gavrilov, A., Kravtsov, S., Buyanova, M., Mukhin, D., Loskutov, E., and Feigin, A.: Forced response and internal variability in ensem-
bles of climate simulations: identification and analysis using linear dynamical mode decomposition, *Climate Dynamics*, 62, 1783–1810,
<https://doi.org/10.1007/S00382-023-06995-1>, 2024.



- 870 Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., Ohgaito, R., Ito, A., Yamazaki, D., Okajima, H., Ito, A., Takata, K., Ogochi, K., Watanabe, S., and Kawamiya, M.: Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks, *Geoscientific Model Development*, 13, 2197–2244, <https://doi.org/10.5194/gmd-13-2197-2020>, 2020.
- Henckel, L., Würtzen, T., and Weichwald, S.: Adjustment Identification Distance: A gadjid for Causal Structure Learning, in: The 40th
875 Conference on Uncertainty in Artificial Intelligence, <https://proceedings.mlr.press/v244/henckel24a.html>, 2024.
- Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M.: Selecting a climate model subset to optimise key ensemble properties, *Earth System Dynamics*, 9, 135–151, <https://doi.org/10.5194/esd-9-135-2018>, 2018.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren,
880 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.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, *Geophysical Research Letters*, 40, 1194–1199, <https://doi.org/10.1002/grl.50256>, 2013.
885
- Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, *Geophysical Research Letters*, 44, 1909–1918, <https://doi.org/10.1002/2016GL072012>, 2017.
- Kravtsov, S., Gavrillov, A., Buyanova, M., Loskutov, E., and Feigin, A.: Forced signal and predictability in a prototype climate model: Implications for fingerprinting based detection in the presence of multidecadal natural variability, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32, 123 130, <https://doi.org/10.1063/5.0106514>, 2022.
890
- Kuma, P., Bender, F. A.-M., and Jönsson, A. R.: Climate Model Code Genealogy and Its Relation to Climate Feedbacks and Sensitivity, *Journal of Advances in Modeling Earth Systems*, 15, e2022MS003 588, <https://doi.org/10.1029/2022MS003588>, e2022MS003588 2022MS003588, 2023.
- 895 Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., and Knutti, R.: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America, *Journal of Geophysical Research: Atmospheres*, 123, 4509–4526, <https://doi.org/10.1029/2017JD027992>, 2018.
- Masson, D. and Knutti, R.: Climate model genealogy, *Geophysical Research Letters*, 38, L08 703, <https://doi.org/10.1029/2011GL046864>, 2011.
- 900 Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenez-de-la Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S.-S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick,
905 C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., von Storch, J.-S., Tian, F., Voigt, A., Vértes, P., Wieners, K.-H., Wilkenskjaeld, S., Winkler, A., and Roeckner, E.: Developments in the MPI-M



- Earth System Model version 1.2 (MPI-ESM1.2) and its response to increasing CO₂, *Journal of Advances in Modeling Earth Systems*, 11, 998–1038, <https://doi.org/10.1029/2018MS001400>, 2019.
- 910 Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., and Knutti, R.: An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles, *Earth System Dynamics*, 11, 807–834, <https://doi.org/10.5194/esd-11-807-2020>, 2020.
- Merrifield, A. L., Brunner, L., Lorenz, R., Humphrey, V., and Knutti, R.: Climate model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for regional applications, *Geoscientific Model Development*, 16, 4715–4747, <https://doi.org/10.5194/gmd-16-4715-2023>, 2023.
- 915 Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F., and Marotzke, J.: A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR), *Journal of Advances in Modeling Earth Systems*, 10, 1383–1413, <https://doi.org/10.1029/2017MS001217>, 2018.
- Murtagh, F. and Contreras, P.: Algorithms for hierarchical clustering: an overview, *WIREs Data Mining and Knowledge Discovery*, 2, 86–97, <https://doi.org/10.1002/widm.53>, 2012.
- 920 Nowack, P., Runge, J., Eyring, V., and Haigh, J. D.: Causal networks for climate model evaluation and constrained projections, *Nature Communications*, 11, 1415, <https://doi.org/10.1038/s41467-020-15195-y>, 2020.
- O'Reilly, C. H., Brunner, L., Qasmi, S., Nogherotto, R., Ballinger, A. P., Booth, B., Befort, D. J., Knutti, R., Schurer, A. P., Ribes, A., Weisheimer, A., Coppola, E., and McSweeney, C.: Assessing observational constraints on future European climate in an out-of-sample framework, *npj Climate and Atmospheric Science*, 7, 95, <https://doi.org/10.1038/s41612-024-00648-8>, 2024.
- 925 Peters, J. and Bühlmann, P.: Structural intervention distance for evaluating causal graphs, *Neural Comput.*, 27, 771–799, https://doi.org/10.1162/NECO_a_00708, 2015.
- 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.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview, *Geoscientific Model Development*, 13, 1179–1199, [https://doi.org/10.5194/gmd-13-](https://doi.org/10.5194/gmd-13-1179-2020)
- 930 1179-2020, 2020.
- Runge, J., Bathiany, S., Boltt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J.: Inferring causation from time series in Earth system sciences, *Nature Communications*, 10, 2553, <https://doi.org/10.1038/s41467-019-10105-3>, 2019a.
- 935 Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D.: Detecting and quantifying causal associations in large nonlinear time series datasets, *Science Advances*, 5, eaau4996, <https://doi.org/10.1126/sciadv.aau4996>, 2019b.
- Runge, J., Gerhardus, A., Varando, G., Eyring, V., and Camps-Valls, G.: Causal inference for time series, *Nature Reviews Earth and Environment*, 4, 487–505, <https://doi.org/10.1038/s43017-023-00431-y>, 2023.
- Sanderson, B. M., Knutti, R., and Caldwell, P.: A representative democracy to reduce interdependency in a multimodel ensemble, *Journal of*
- 940 *Climate*, 28, 5171–5194, <https://doi.org/10.1175/JCLI-D-14-00362.1>, 2015.
- Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, *Geoscientific Model Development*, 10, 2379–2395, <https://doi.org/10.5194/gmd-10-2379-2017>, 2017.
- Seabold, S. and Perktold, J.: statsmodels: Econometric and statistical modeling with Python, in: 9th Python in Science Conference, <https://doi.org/10.25080/Majora-92bf1922-011>, 2010.



- 945 Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., Decharme, B., Delire, C., Berthet, S., Chevallier, M., Sénési, S., Franchistéguy, L., Vial, J., Mallet, M., Joetzjer, E., Geoffroy, O., Guérémy, J.-F., Moine, M.-P., Msadek, R., Ribes, A., Rocher, M., Roehrig, R., Salas y Mélia, D., Sanchez, E., Terray, L., Valcke, S., Waldman, R., Aumont, O., Bopp, L., Deshayes, J., Éthé, C., and Madec, G.: Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate, *Journal of Advances in Modeling Earth Systems*, 11, 4182–4227, <https://doi.org/10.1029/2019MS001791>, 2019.
- 950 Seland, Ø., Bentsen, M., Olivie, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y.-C., 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., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T., and Schulz, M.: Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations, *Geoscientific Model Development*, 13, 6165–6200, <https://doi.org/10.5194/gmd-13-6165-2020>, 2020.
- 955 Shepherd, T.: Atmospheric circulation as a source of uncertainty in climate change projections, *Nature Geoscience*, 7, 703–708, <https://doi.org/10.1038/NGEO2253>, 2014.
- Shin, Y., Lee, Y., and Park, J.-S.: A weighting scheme in a multi-model ensemble for bias-corrected climate simulation, *Atmosphere*, 11, 775, <https://doi.org/10.3390/atmos11080775>, 2020.
- Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizuki, T., Yoshimura, K., Takata, K., O’ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M., and Kimoto, M.: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6, *Geoscientific Model Development*, 12, 2727–2765, <https://doi.org/10.5194/gmd-12-2727-2019>, 2019.
- Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365, 2053–2075, <https://doi.org/10.1098/rsta.2007.2076>, 2007.
- 965 Tsamardinos, I., Brown, L. E., and Aliferis, C. F.: The max-min hill-climbing Bayesian network structure learning algorithm, *Machine Learning*, 65, 31–78, <https://doi.org/10.1007/s10994-006-6889-7>, 2006.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., and van Mulbregt, P.: SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nature Methods*, 17, 261–272, <https://doi.org/10.1038/s41592-019-0686-2>, 2020.
- 970 Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., Nabat, P., Roehrig, R., Salas y Mélia, D., Séférian, R., Valcke, S., Beau, I., Belamari, S., Berthet, S., Cassou, C., Cattiaux, J., Deshayes, J., Douville, H., Ethe, C., Franchistéguy, L., Geoffroy, O., Lévy, C., Madec, G., Méléard, J., Mignot, J., Miller, J., Moine, M.-P., Nabat, P., Roehrig, R., Salas y Mélia, D., Séférian, R., Valcke, S., and Waldman, R.: Evaluation of CMIP6 DECK experiments with CNRM-CM6-1, *Journal of Advances in Modeling Earth Systems*, 11, 2177–2213, <https://doi.org/10.1029/2019MS001683>, 2019.
- Walters, D., Baran, A. J., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., Furtado, K., Hill, P., Lock, A., Mannes, J., Morcrette, C., Mulcahy, J., Sanchez, C., Smith, C., Stratton, R., Tennant, W., Tomassini, L., Van Weverberg, K., Vosper, S., Willett, M., Browse, J., Bushell, A., Carslaw, K., Dalvi, M., Essery, R., Gedney, N., Hardiman, S., Johnson, B., Johnson, C., Jones, A., Jones, C., Mann, G., Milton, S., Rumbold, H., Sellar, A., Ujiie, M., Whitall, M., Williams, K., and Zerroukat, M.: The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations, *Geoscientific Model Development*, 12, 1909–1963, <https://doi.org/10.5194/gmd-12-1909-2019>, 2019.
- 980



- Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of model weighting in multimodel climate projections, *Journal of Climate*, 23, 4175–4191, <https://doi.org/10.1175/2010JCLI3594.1>, 2010.
- 985 Youden, W. J.: Index for rating diagnostic tests, *Cancer*, 3, 32–35, [https://doi.org/10.1002/1097-0142\(1950\)3:1<32::AID-CNCR2820030106>3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3), 1950.
- Ziehn, T., Chamberlain, M. A., Law, R. M., Lenton, A., Bodman, R. W., Dix, M., Stevens, L., Wang, Y.-P., and Srbinovsky, J.: The Australian Earth System Model: ACCESS-ESM1.5, *Journal of Southern Hemisphere Earth Systems Science*, 70, 193–214, <https://doi.org/10.1071/ES19035>, 2020.