

Using decadal forecasts of modes of variability to refine the first few years of climate projections

David Sexton

david.sexton@metoffice.gov.uk

Met Office <https://orcid.org/0000-0002-0004-0518>

Jason Lowe

Met Office

Carol McSweeney

Met Office

Research Article

Keywords: Merging climate information, Seamless information, Climate projections, Initialised decadal forecasts, Model weighting, Modes of variability

Posted Date: August 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9995238/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Title: Using decadal forecasts of modes of variability to refine the first few years of climate projections

David M. H. Sexton¹, Jason A. Lowe¹, and Carol F. McSweeney¹

5 ¹ Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK

Corresponding author: David Sexton (david.sexton@metoffice.gov.uk)

ORCID Numbers:

David Sexton	0000-0002-0004-0518
Jason Lowe	0000-0002-8201-3926
Carol McSweeney	0000-0001-8898-2093

10

15 **Acknowledgements** This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081460. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This work was also funded by U.K. Research and Innovation under the

20 U.K. government's Horizon Europe funding Guarantee (Grant 10060007). We would like to thank Glen Harris for his detailed description of the location of the hindcasts not available from the CMIP6 DCP archive. We would like to thank Markus Donat for his very helpful and constructive feedback on the initial concept, and useful discussions with Doug Smith, Antje Weisheimer, James Murphy, Beena Balan Sarojini and Adnan Abid.

25

Abstract

Adaptation planners increasingly use climate projections, sometimes weighted or filtered by their ability to represent average climate and recent temperature trends. Decadal forecasts initialised by the current state of the climate system offer greater near-term accuracy. However, integrating both sources of information coherently across time scales is challenging. Here, we complement existing methods by weighting projections based on their near-term evolution. This incorporates the value of initialization without directly using the forecasts enabling a seamless transition to longer term projections. Our method estimates a weight for each projection after removing the forced signal from both the forecasts and the long-term projections, so that they are independent of previously applied constraints. We use skill and reliability metrics to show how well the “merged product” captures the skill-based information of the initialised forecasts at short lead times and the reliability-based information at longer time scales. For global temperature, ENSO, and North Atlantic subpolar gyre temperatures, the merged product adds value over raw projections for lead times of 2, 3 and 5 years respectively. We provide guidance to decision makers on the appropriate lead times for using forecasts, projections, or the merged product. For the winter North Atlantic Oscillation, consistent with previous studies, biases in both signal and variability lead to issues with both reliability and usable skill; these problems need to be resolved for affected variables before any merging is applied.

Keywords: Merging climate information; Seamless information; Climate projections; Initialised decadal forecasts; Model weighting; Modes of variability.

1. Introduction

Climate information used for adaptation planning on time horizons of decades is often based on a range of possible outcomes estimated from sets of long-term simulations from different climate models (Intergovernmental Panel on Climate Change (IPCC) 2023). For a given set of historical and future anthropogenic and natural forcings, the spread in these projections derives from two sources: the epistemic uncertainty in the forced response from the different models to the realistic forcing, and internal variability which is sampled independently from the real world. Adaptation decisions might aim for resilience across the plausible range of outcomes, or at least stress test against an upper bound of the projections (Climate Financial Risk Forum (CFRF) 2024). Climate projections do not attempt to assimilate the current state of the climate system, which is appropriate for longer term decision beyond a decade or more, but misses potential information on how the recent state of the climate system influences the near future. For decisions at short lead times, planners may test their decisions against the projections or instead use present day climatology (Thompson et al. 2017). An alternative source of climate information for near-term planning is decadal forecasts (Intergovernmental Panel on Climate Change (IPCC) 2014), which are initialised using the observed oceanic (and sometimes atmospheric) state to constrain the predictable component of the internal variability. For some hazards of interest, retrospective forecasts (hindcasts) show varying degrees of increased precision that might lead to improved decision making. The current uptake of using this decadal information is limited but a range of projects are attempting to work with potential users to understand and demonstrate how they might be used (e.g. see <https://www.aspect-project.eu/>).

For decision makers planning across time scales from a year ahead to several decades, it is natural to seek coherent information between decadal forecasts and projections. A recent survey of users of the UK's national climate projections (Mittal et al. 2025) revealed some sectors make decisions that bridge these time scales (insurance, hydrology, agriculture, nature management). During a focus group meeting in that survey, some participants expressed a need to include the added value from initialisation when the forecasts demonstrate skill and deviate appreciably from the projected trend. Taking an overall view of physical climate risk in private sector or governmental organisations requires a coherent assessment across timescales, but this is currently elusive. We propose an approach that could aid this risk assessment.

Several challenges make using decadal forecasts and projections difficult. Decadal forecasts are more complex; they require adjustments for drift, and some models are affected by

spurious “initialisation shocks”, both of which are typically unfamiliar to those who only focus on using projections (Delgado-Torres et al. 2025). Even if these issues are mitigated, inconsistencies remain across the data sets, such as differing forcing scenarios or climate models. Finally, one must handle differences between the precision-based information offered by decadal forecasts versus the projected range of plausible outcomes. As a result, many users of climate information prefer to rely on a single, familiar source of climate information. In this paper, we focus on developing a “merged product” based on the projections that incorporates the added value of initialisation in the near-term and enables the possibility of consistent information across time scales.

For the merged product to be useful for real-world decision making and planning, it needs to successfully bridge between two very different decision-making paradigms, each with its own philosophy and performance metrics. The value of initialisation is measured by the skill of the hindcasts, primarily the correlation with observations. However, a significant part of the added value comes from constraining the variability; this “Contribution from initialisation” is measured by the correlation of residual variations after removing the forced signal from the hindcasts and observations (Smith et al. 2019). Studies also show other benefits, such as improvements to the forced signal (Doblas-Reyes et al. 2013) and initialised forecasts provide other valuable information such as insights into mechanisms (Patrizio et al. 2025) or data sets of unprecedented or unrealised present-day events (Thompson et al. 2017). In contrast, the long-term projections cannot be verified against observations so much care is taken to guard against over-confident decisions that could result from over-constrained projections. Two complementary approaches to avoid over-constraining the projections are: (i) using multivariate constraints to prevent overly rewarding a climate model for the wrong reasons (Sexton et al. 2012), and (ii) accounting for structural error, which weakens the likelihood constraint in probabilistic projections (Goldstein and Rougier 2004). As a result, projections are often intentionally over-dispersive, as demonstrated by reliability metrics such as spread-over-error or the fraction of times the observed values fall within the 80% credible interval (Murphy et al. 2025). Here, we measure the performance of the merged product for various large-scale indices aiming for both reliability and skill that is comparable to the hindcasts at short lead times but reducing over time as predictability decreases. For variables where the merged product performs well, planners will have three sources of climate information to consider: forecasts, long-term projections and the merged product. These metrics will guide the choice of which one to use.

Several approaches already exist for temporally merging initialised forecasts with a set of unconstrained long-term projections. One approach, known as ‘shadowing’, selects a subset of realisations from the long-term projections that best match the decadal forecast e.g. according to its fit to a large-scale index which is an important driver of regional climate (Befort et al. 2020), or to a regional SST pattern (Delgado-Torres et al. 2025). In both cases, the size of the retained subset is pre-defined to e.g. 20-50, and used for longer time scales. This is effectively using the initial conditions to constrain long-term projections beyond the time scales of predictability. The relatively small sample size is potentially under-dispersive, making it less suitable for resilience-based decisions that need to consider the full range of plausible outcomes; it is also problematic if rare events are of interest. The current shadowing methods only apply to an unfiltered set of projections. They do not account for the situation where the projections are already weighted by their goodness of fit to mean climate and/or recent large-scale temperature trends (Harris et al. 2013; Brunner et al. 2020). There is an important extra consideration when including additional constraints, which is to avoid double-counting of information (or properly account for any interdependence between constraints). This mitigates against over-confident information that could adversely affect the resilience-based decision-making. We avoid double-counting by removing the long-term forced signal in the forecast and focusing on the constraint on variability. Specifically, we match deviations about the forced signal in the long-term projections to the variations about the uninitialised forced signal in the ensemble forecast. This approach is independent of any weights based on the forced signal in the long-term projections. It differs from Befort et al’s shadowing approach in two ways. First, it protects against projections being retained by having compensating errors between the forced signal and the internal variability. Second, the weights relax with lead time, so that the information from the current climate state aims to not over-constrain the projections beyond the time scales of predictability.

In section 2, we describe the data sources for the projections, decadal hindcasts and reanalysis used to estimate performance metrics. We restrict ourselves to available mainstream data sets for projections and forecasts that have not undergone prior bias adjustment other than the standard drift correction. As a result, these data sets may still exhibit biases in variability and signal. This means that the merged product may inherit such issues, but addressing these is beyond the scope of this work. Section 3 introduces a conceptual model relating uninitialised projections and initialised forecasts based on true predictable signals. We then describe how to estimate the terms in the conceptual model, the likelihood weights, and how the skill and reliability metrics account for this weighting. We illustrate the various steps for summer

subpolar North Atlantic surface temperatures. In section 4, we apply the method to this index and three additional three large-scale indices to explore a range of levels of skill and reliability. We assess the merging using skill and reliability metrics. Section 5 summarise the findings followed by a discussion in section 6 on implications for users and directions for further research and development.

2. Data

2.1 Data sources

We use twelve decadal prediction systems (see Table 1). All systems initialise the global ocean using observations, consisting in most cases of three-dimensional fields of temperature and salinity. Most systems also initialise the atmosphere and some include direct initialisation of sea ice and land surface components. Each institute provides several realisations, typically 10, which differ in terms of their initial conditions. We use ten-year hindcasts from every year from 1963 to 2016, initialised late in the year. External forcing follows the CMIP6 specification of observed changes to 2014 (Eyring et al. 2016), switching subsequently to the SSP2-4.5 scenario (O'Neill et al. 2016).

To evaluate the skill and reliability metrics for the decadal hindcasts and for the merged product, we used ERA5 reanalysis (Hersbach et al. 2020). For the projections, we use all available realisations from 48 CMIP6 models produced at 23 different worldwide institutions (see Table 2). For some models, there is only one realisation. Beyond 2004, the high emissions SSP5-85 scenario is used due to its large number of available realisations and to demonstrate the method working for a scenario that is different to that used in the decadal forecasts.

We test the method for four temperature or circulation related variables (defined in Table 3). All available forecast and projection realisations were used for each variable. Fig. 1 shows that they range from high levels of skill for ENSO and annual global mean temperature at lead time T+1 as measured by the residual anomaly correlation, to modest skill for the winter North Atlantic circulation. These indices are also sample a range of levels of predictability from a few years for the North Atlantic temperatures to one year for winter North Atlantic circulation. Note CanESM5 is not used for summer SPNA due to its much lower skill than the other models; this is due to a known initialisation shock.

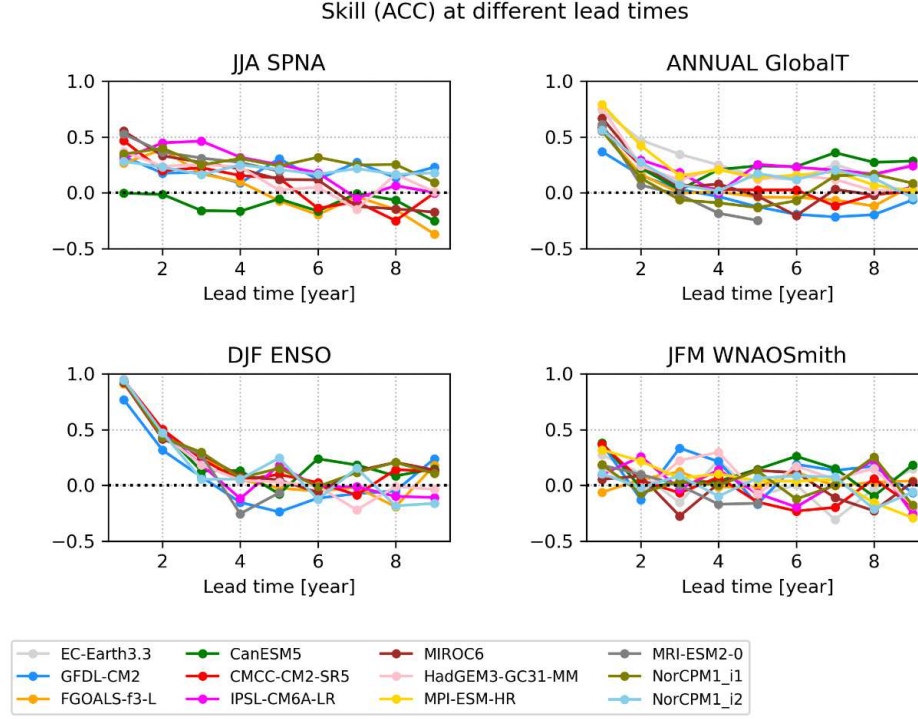


Fig. 1 Residual correlation for decadal forecasts of JJA SPNA. The correlation is estimated using the approach in section 3. This is an adaptation of the original method in Smith et al (2019), with a different way to remove the forced signal

185

3. Methods

3.1 Conceptual model for shadowing

The aim of shadowing is to post-process a set of projections over time, t , of a quantity of interest based on the level of similarity to a set of initialised forecasts from start time, S . The measure of similarity is typically correlation-based but it can be mean square error-based as used here; both incorporate phase or pattern information, but the mean square errors also includes amplitude information. We assume the quantity is univariate, but our approach could easily be generalised to be multivariate.

The projection data, y_{mrt} , is a pooled set of time series from $m = 1, \dots, M$ climate models, each with $r = 1, \dots, R_m$ realisations. This structure reflects the design of the CMIP6 multimodel data set but also covers probabilistic projections, where M surrogate models may be emulated with multiple realisations. As in conceptual models used for the Analysis of Variance of initial condition ensembles, the projection data can be expressed as the sum of the true signal, $\mu_{mt} + v_{mt}$ (split into long and short time scales respectively), and internal variability, ε_{mrt} ,

$$y_{mrt} = \mu_{mt} + v_{mt} + \varepsilon_{mrt}. \quad (\text{Eqn. 1})$$

The forecast data, $x_{ij\tau}$, initialised from start time S at lead time, τ (which varies between 1 and τ_{max}) is based on a pooled set of time series from $i = 1, \dots, I$ climate models (possibly with some overlap with those used for the projections), each with $j = 1, \dots, J_i$ realisations. This can be decomposed into several terms:

$$x_{ijS\tau} = d_{iS\tau} + \delta_{iS\tau} + (\beta\mu_{it} + \gamma v_{it} + \Delta_{iS\tau}) + \eta_{ijS\tau}, \quad (\text{Eqn. 2})$$

where $d_{iS\tau}$ is the true drift, $\delta_{iS\tau}$ is the true initialisation shock, and $\eta_{ijS\tau}$ is the internal variability for the j th realisation, at time $t = S + \tau$. There is also the true predictable signal, the actual value which we can only estimate from the data. This has been separated further into (i) the true forced signal at short and long time scales, $\beta\mu_{it} + \gamma v_{it}$, which is linearly related from the uninitialised model simulation at time t (noting that an actual uninitialised simulation may not always exist), and (ii) the true residual predictable signal, identified as the contribution from initialisation, $\Delta_{iS\tau}$. The first term could be assumed equal to the signal from the uninitialised runs, but we allow for a linear relationship to reflect the original method used to estimate the residual anomaly correlation.

The aim of our shadowing approach is to weight the projections for $t = S + 1, S + 2, \dots, S + \tau_{max}$ according to how well the internal variability in the projections matches the ensemble mean and spread of $\Delta_{iS\tau}$. This is accomplished in two steps. First, likelihood weights are estimated separately for each lead time, $\tau = 1, \dots, \tau_{max}$, using the average of the pooled forecasts from the lead time $\tau, \dots, \tau_{max}$. So, the initial weights for lead time τ do not use the forecast prior to that lead time. As both the forecast signal weakens and the spread of the forecasts increases with lead time, the weights will flatten out across the pooled set of projections tending towards $1/M$. Second, to avoid rapid jumps the weights are smoothed across the lead times to make the weights transition more coherently from one lead time to the next. This second step recognises the waning influence of initialisation as predictability decreases over time. We chose to smooth by using the maximum up to lead time τ , thereby retaining information about any initial predictability. This might not be an optimal choice and warrants further research, but tests showed only minor sensitivity to other choices such as running averages and geometrical means.

3.2 Estimation of terms in conceptual model

We estimate the terms in the idealised conceptual model, recognising that the real-world conditions are different and that certain assumptions and pragmatic steps are necessary. We are also restricted to using mainstream approaches which are applied to projections and decadal forecasts. In section 6, we discuss opportunities to further develop these methods.

a) Estimating signal and noise in CMIP6 projections

The forced signal from a set of R realisations for a given climate model has typically been estimated as the ensemble mean. The variance of this ensemble mean compared to the variance of the true forced signal is biased high by $1/R$ of the variance of internal variability. To accommodate climate models with only a single or a few realizations, we further reduce the variance of the estimated signal by applying a 32-year low-pass filter to each realisation, recognizing that the anthropogenically forced signal evolves on longer timescales than year-to-year variability. We chose 32 years as it was safely longer than the period of any variations in deviations from the ensemble averages. For those climate models with multiple realisations, we then take the ensemble average of the smoothed time series and estimate an initial set of deviations from the ensemble mean (see Fig. 2). For single realisations, we simply remove the low pass filtered signal.

The filtering does not remove forced signals on shorter time scales much less than 32 years such as those due to solar activity and volcanic eruptions. In doing so, we have assumed the short-term forcing signal is a common shape for all projections (effectively setting v_{it} to v_t in Eqn. 2). We estimate the contribution from these forcings by averaging the pooled set of deviations and subtracting this average to obtain the final set of deviations. As shown by the red line in Supp Fig. S1, this term is only significant for the annual global mean temperature among our indices. It is the final set, $\hat{\epsilon}_{mrt}$ that are compared against the forecast to estimate the likelihood weights for shadowing. Supp Fig. S2 shows that the spread of the deviations is consistent with deviations in the reanalysis data but that there are a few models, notably ACCESS-ESM1-5 and ACCESS-CM2 that have too much high frequency variability for summer SPNA.

Some of the ensemble means of the climate models with multiple realisations, and the high frequency forced signal will also be used to estimate the contribution from initialisation.

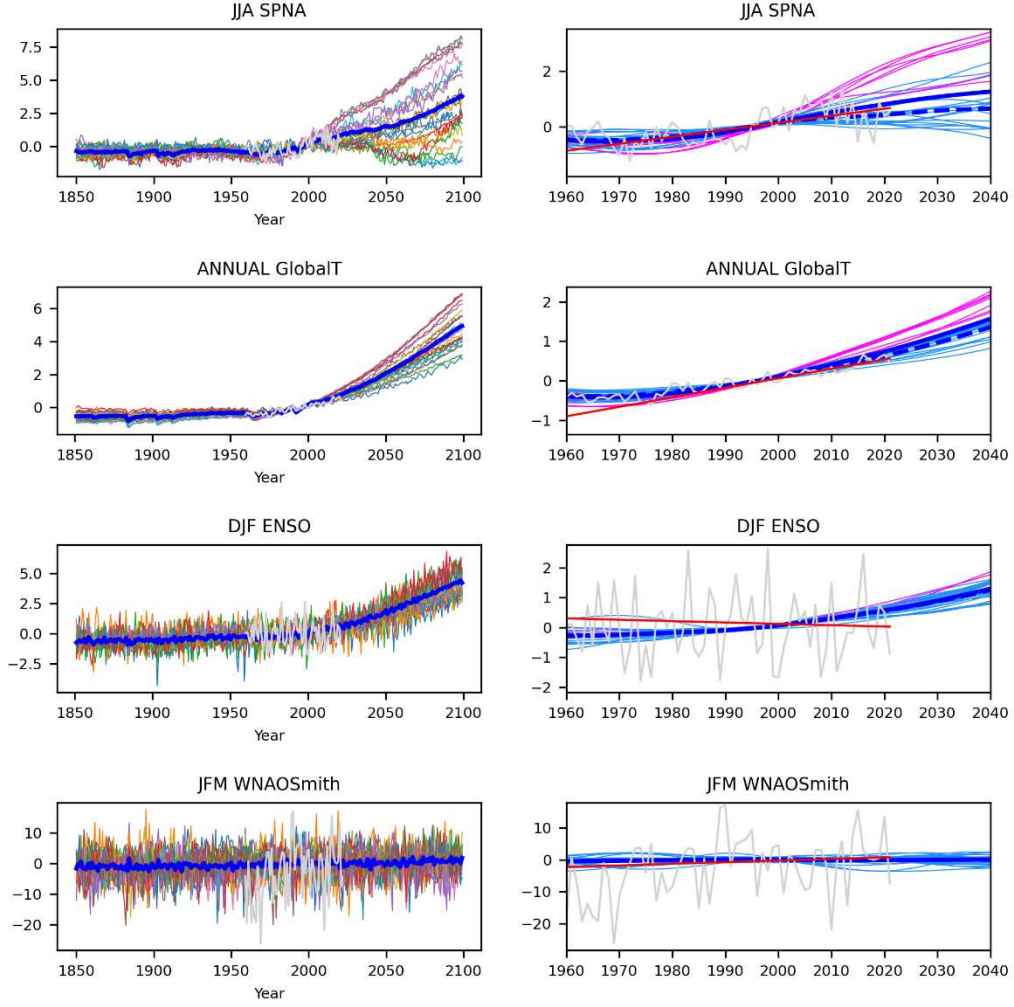


Fig.2 The individual realisations in the CMIP6 projections (thin lines) compared to the reanalysis time series (light grey) and the average over the pooled set (dark blue) for each large-scale index (each row). The left panels show unfilted time series. In the right panels, only the low-pass filtered averages are shown for each ensemble. The magenta lines show ensembles that are not consistent with the reanalysis trend, otherwise the lines are light blue. The red line shows the trend line estimated for the reanalysis time series from 1991-2020, extrapolated back to 1960

b) Estimating contribution from initialisation in decadal forecasts

In practice, the forecast data,

$$x_{ijs\tau}^{BC} = \beta\mu_{it} + \gamma v_{it} + \Delta_{is\tau} + \eta_{ijs\tau}, \quad (\text{Eqn. 3}),$$

is drift-corrected using information from the hindcast data set based on the method described in Boer et al. (2016). This removes a constant term as a function of lead time only, but Kharin et

al. (2012) have shown that some forecast models exhibit drifts that are also dependent on the start date, S (notably those models with relatively high climate sensitivities which causes their historical simulations to warm too rapidly). The initialisation shock only affects some models enough to deteriorate the quality of the forecast and is typically consistent across realisations.

280 Here we assume that the drift adjustment works and there is no initialisation shock, recognising that if this is not the case, our estimates below will be contaminated by these. The projections have their own biases, but they do not suffer from these two artefacts of the initialisation. Therefore, the outcome from the shadowing application is also free of these two sources of error.

285 Smith et al. (2019) estimate the contribution of initialisation by regressing out the multimodel mean forced signal. The use of the multimodel mean is problematic because (i) ensemble averages across different climate models can vary widely, indicating an appreciable degree of signal uncertainty (see Fig. 2); (ii) cancellation across this broad range of signals can result in an average with relatively low amplitude that may not be representative of the signal in the real
290 system (Smith et al. 2025). To estimate the high frequency signal (see Fig. 3), we remove the “forced contribution” for each forecast model (grey lines) individually by regressing out the average of a suitably chosen ensemble within the projections set (red lines). For forecast models that have a corresponding uninitialised projection, we use that ensemble mean. For other forecast models, we use the ensemble mean that most strongly correlates with the
295 hindcasts with the longest available lead time ($T+10$ or $T+9$) where the uninitialised forced signal is most dominant. This regression is represented by the β and γ terms in Eqns. 2 and 3.

JJA SPNA T+1

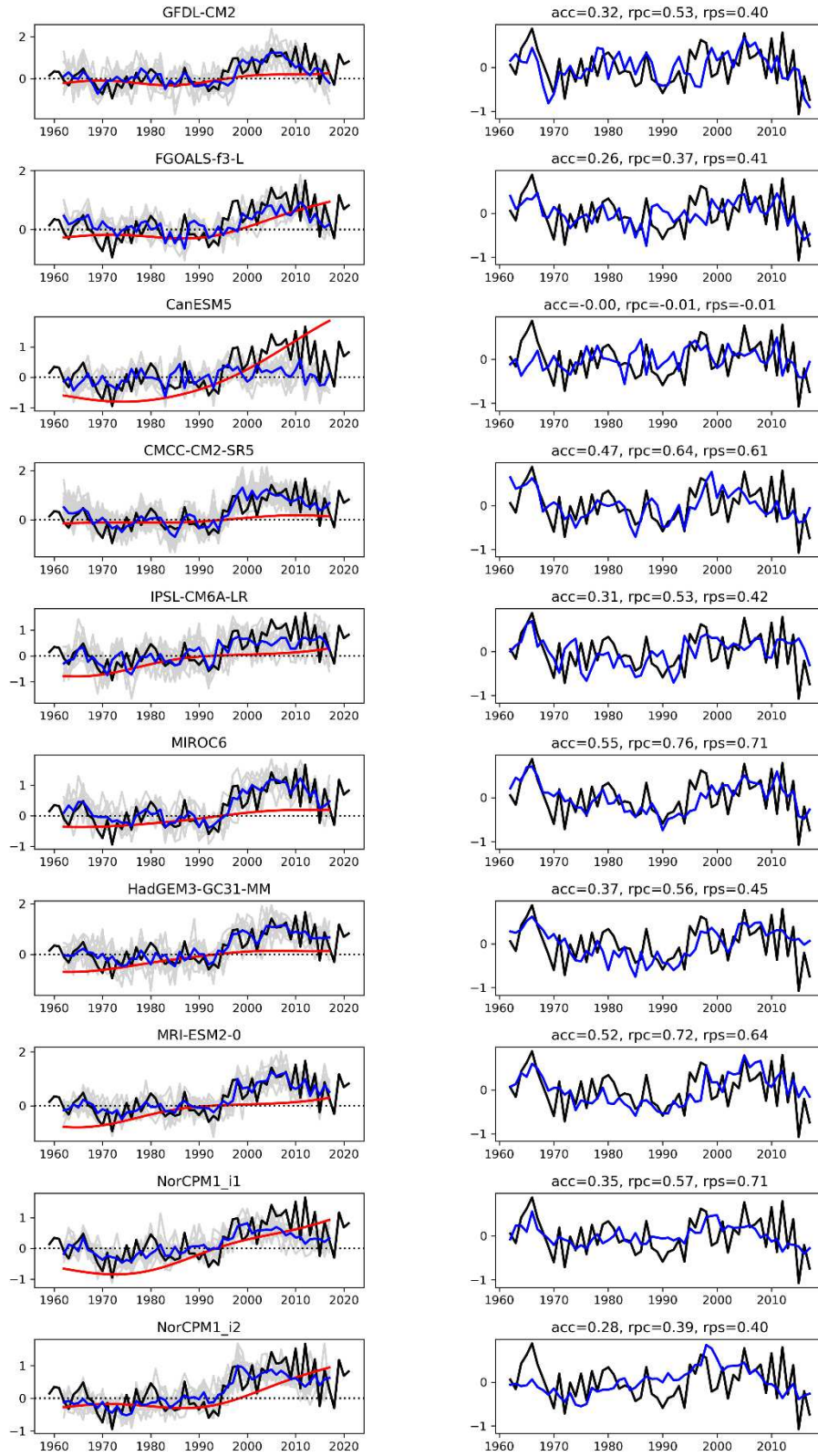


Fig.3 Estimation of the ensemble mean residual signal for each decadal forecast for summer SPNA. (left panels) The drift-corrected anomaly time series for each realisation of the forecast

300 (grey lines) and their ensemble average (blue line) compared to the smooth estimated forced signal from the CMIP6 projections appropriate for that forecast model (red line) and the reanalysis (black line). (right panels) Comparison of the ensemble mean residuals with the reanalysis residuals after the uninitialised short- and long-term signals are regressed out

305 c) Estimating simple trend weights for each CMIP6 ensemble

Our approach is designed to work for situations where the projections have already been weighted by how well each simulation captures forced historical trends. These weighting schemes are often complex and may also consider other factors, such as their ability to represent observed climate averages (Harris et al. 2013; Brunner et al. 2020). To focus on the additional weighting applied for shadowing, we adopt a simple trend weighting. This weighting is likely to be too strong as it is univariate and does not account for structural error. First, we estimate the trend and its error variance (assuming red noise) from the reanalysis trend of the quantity of interest for the period 1981-2020. For each realisation in the pooled multimodel set, we estimate its trend over the same period and determine the likelihood of that trend given the mean and error variance of the observed trend. We then normalise these trend weights. The magenta lines in Fig. 2 are indicative of ensemble averages that are not consistent with the reanalysis trend.

d) Estimating likelihood weights for shadowing

For a given lead time, τ , and for a given realisation of a projection (which specifies m and r), we estimate the likelihood of the T -length vector, $\hat{\epsilon}_{mrt}$ where $T = \tau_{max} - \tau + 1$ and $t = S + \tau, \dots, S + \tau_{max}$, from the T -dimensional multivariate Gaussian distribution, $\mathcal{N}(\overline{\Delta_{S\tau}}, \Sigma_{S\tau}^{mean} + \Sigma_{\tau}^{res})$. The mean term is from the average of the contribution from initialisation of the pooled set of forecast models (blue lines in Fig. 3 right panels). There are two contributions to the variance. The first term is the variance of the average contributions from initialisation across the forecast models (see Supp Fig. S3).

The second term accounts for the uncertainty in the projection term. If we had an infinite sized forecast, even with an imperfect model, we would have a precise, zero variance, estimate of the forecast signal. We would still have to account for the possibility that a residual variation in the projection set that matches this perfect signal exactly, might have done so for fortuitous reasons with signal being offset by internal variability. We estimate this term from the deviations

of the individual forecast realisations from their estimated contribution of initialisation, the residuals of the residuals.

3.3 Metrics for evaluating performance

335 To evaluate the performance of the shadow-weighted projections, we apply the weighting to the start dates covered by the decadal hindcasts and compare against the observations. We also do this for the hindcasts and unweighted projections to provide benchmarks. We evaluate three metrics a)-c). The correlation measures skill for precision-based decision-making and we hope this will be similar to the decadal forecasts at short lead times. As outlined in section 1, the
340 other two measure reliability relevant to resilience-based decision-making. By measuring This is done for each lead time, T+1 to T+9 years.

a) Anomaly correlation coefficient (ACC)

The centred Pearson correlation coefficient of the weighted mean of the pooled set of projections.

345 b) Spread-over-error (SOE)

The weighted ensemble standard deviation of the pooled set of projections divided by the root mean square error of the weighted mean of the pooled set of projections. Ideally the value is 1, with less (greater) than one indicating under(over)-dispersion.

c) Reliability of 80% confidence interval (FRAC80)

350 First the 10th and 90th percentiles are estimated from the projections using the weights (uses quantile() method in numpy) for each year of the hindcast set. The metric is the fraction of the hindcast years for which the observed value falls within the 10-90th confidence interval. A reliable forecast would give 0.8.

For the unweighted projection set, the same calculations are done with the weights set to 1/M.

355 For the hindcasts, the metrics are based on the equally weighted set of forecast realisations with no accounting for some institutions having more realisations than others (as used in Murphy et al. (2025)).

We use a final diagnostic, the effective ensemble size (Rougier and Sexton 2007),

$$[\sum_{i=1}^M w_i^2]^{-1}$$

to measure the amount of independent information in an ensemble when the i th member is given a weight w_i which are normalised to sum to 1. For an equally weighted ensemble of size M , the effective ensemble size is M . When some members are downweighted, the effective ensemble size decreases. An ensemble where all members except one are given zero weight has an effective ensemble size of 1.

4. Results

a) Summer subpolar North Atlantic SSTs

Constraining by T+1-9, T+2-9, ..., T+9 (smoothed weights by max)

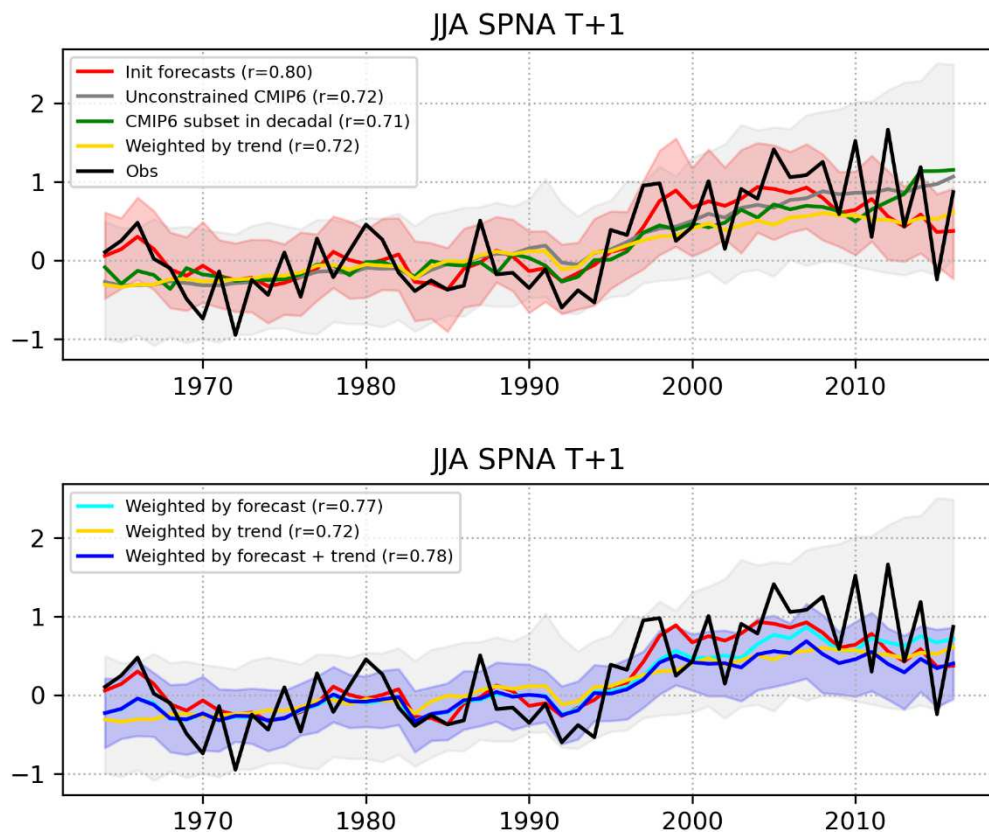


Fig.4 For summer subpolar North Atlantic temperatures over 1963-2016, a comparison of the merged product (blue line and plume) with the original CMIP6 data (grey line and plume), reanalysis (black line) and the initialised hindcasts (red line and plume). Plumes show the 10th-90th percentile confidence interval. A simple trend weighting is applied to the CMIP6 data beforehand (yellow line) and the effect on the merged product without the trend weighting is also shown (light blue line)

375 The first case study is on the summer subpolar North Atlantic temperature, which has
multidecadal variability and is known to have predictability out to a few years and was used by
Befort et al. (2020). The hindcasts at T+1(including the forced signal) have a higher correlation of
0.8 than any of the sets of uninitialised projections (see Fig. 4 top panel). Comparison with the
subset of CMIP6 projections which have corresponding initialised hindcasts suggests this is
380 because they capture the observed increase in SPNA from the late 1990's and then observed
decrease from 2010 onwards. In contrast, the full CMIP6 ensemble mean (grey line) just warms
steadily over the whole period underestimating the temperatures in 1995-2010. Downweighting
the influence of CMIP6 realisations that do not warm at a similar rate to that observed in 1991-
2020 reduces this warming rate (yellow line) as there are known to be a few climate models that
385 warm too fast. This exacerbates the underestimation of the observed warming from 2005
onwards.

Fig. 5 provides the key information about how well our approach meets the user needs across
the two contrasting decision-making paradigms, although some issues are due to problems
inherent in the underlying data. The added value from weighting according to the ability to
390 match the variability predicted by the initialisation can be seen by comparing the skill of CMIP6
projections at different lead times before (grey line) and after (light blue line) weighting. This
shows initialisation adds most value at T+1 and gradually less value until T+6 after which the
correlations are very similar. Weighting according to the level of fit to the recent trend alone
(yellow line) adds skill, more than the initialisation weighting from T+4 onwards. Combining the
395 two weights (blue line) achieves the best results with skill that is similar to that of the hindcasts
for all lead times. The effect of the combined weighting is not as much as sum of their two
incremental changes relative to the unconstrained CMIP6 value (grey line), implying that the
weighting is not additive. This does not imply the design of an independent set of weights is not
optimal. Rather, the combined weighting reflects a complex interplay between differences in
400 signal and variability among the heavily downweighted realizations and those projections that
receive the highest weights. Here, the combine weighting is partly affected by the variability in
the realisations that are downweighted by their relatively poor fit to recent trends. For example,
for T+1, Fig 4 (lower panel) shows that the gains in skill from the initialisation-only and combined
weighting (blue line) might come from different periods. The combined weighting better
405 captures the decrease in temperatures from 2007 as seen relative to the yellow line. In contrast,
the initialisation-only weighted projections (light blue line) better capture more of the observed
warming from 2000-7 but this is because they contain realisations that warm too fast.

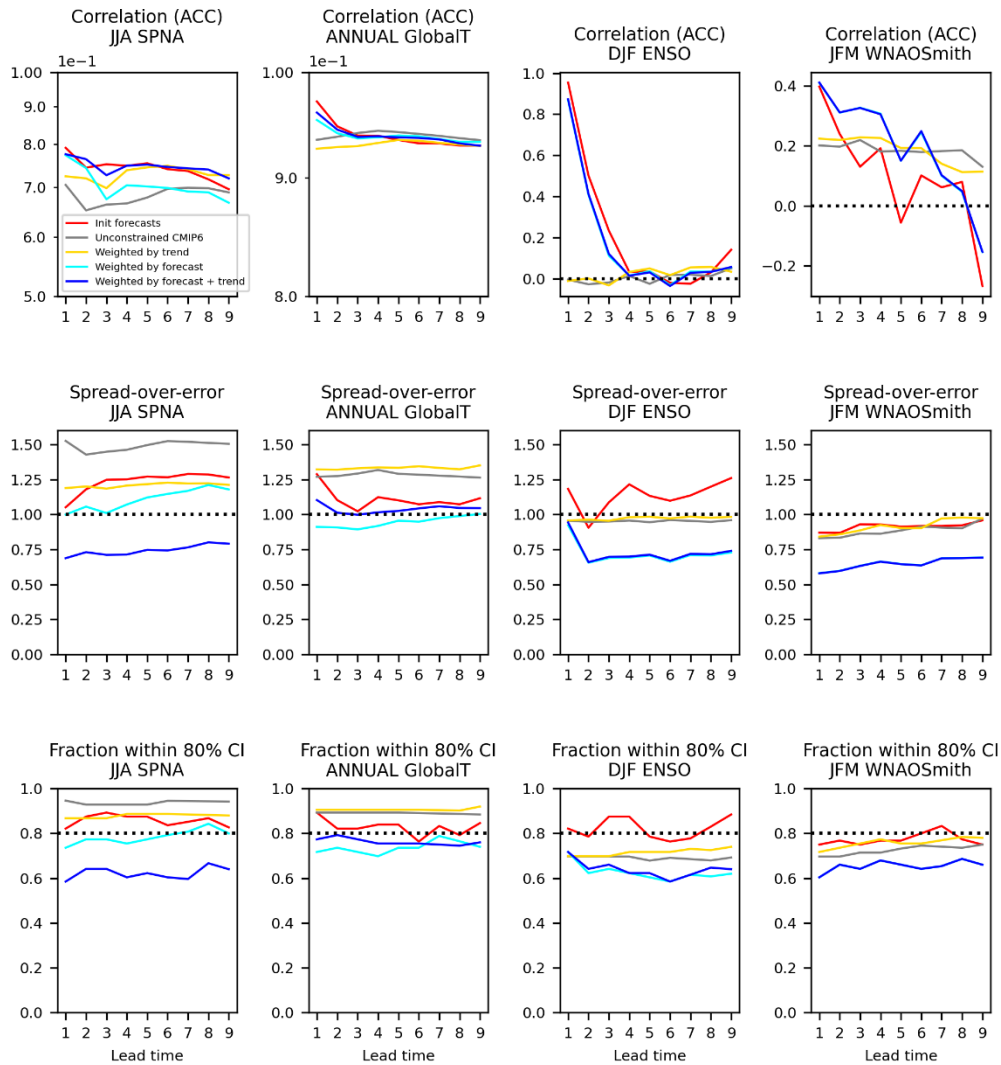


Fig.5 Decision-relevant metrics as a function of lead time for the four large-scale indices (each column). The metrics are anomaly correlation (top row; note that the y-range changes for each index), spread-over-error (middle row), and fraction within the 80% confidence interval (bottom row)

For resilience-based decisions, both reliability metrics indicate that the two sets of CMIP6 projections for summer SPNA are over-dispersive, and that weighting by either the trend or initialized information reduces this dispersion. While our simple trend weighting likely imposes a stronger constraint than more sophisticated schemes, the substantial reduction in dispersion may also result from the loss of the upper part of the bimodal warming distribution in the 2030s. The combined weighting leads to under-dispersion with the observations only falling within the 80% confidence interval 60% of the time. This is mainly because the observed temperatures are

warmer from 1995 onwards and there are a disproportionate number of cases where they exceed the 80th percentile of the combined weighted projections (upper bound of blue plume). There are two main factors for this. First, the hindcasts underestimate the warming in this period suggesting that the signal is too weak in the CMIP6 class of models. Second, the hindcasts show a relatively wide uncertainty (red plume Fig 4 top panel) which would weaken the weighting from initialisation.

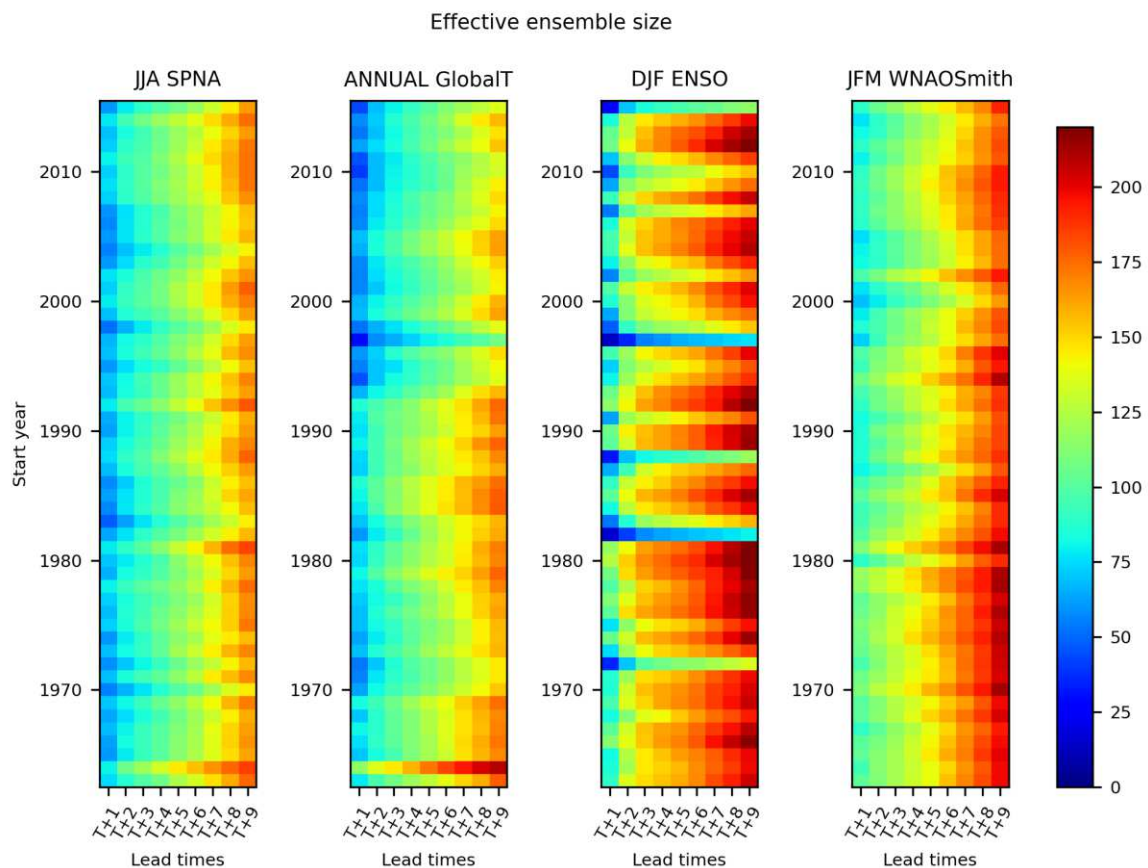


Fig.6 The effective ensemble size as a function of lead time and start year for each of the four large-scale indices

The observed summer SPNA values all lie within the over-dispersive CMIP6 projections (grey plume). For T+1, the effective ensemble size is considerably larger than the subjectively chosen 20 or 30 best realisations used in other shadowing approaches (see Fig. 6). Fig. 6 shows that the effective ensemble size grows with lead time as the added value from the initialisation wanes. By T+9 only about 70% of the ensemble size is effectively being retained after the weighting. This is mainly due to the trend weighting but might also be the elimination of realisations who have poorly represented variability that does not match the observed variability. It also shows that

some start dates have lower effective ensemble size that persists through lead times, suggesting that these are ‘windows of opportunity’ with more predictable situations. However, this might also suggest that weights from early lead times are dominating for later lead times so that the step to smooth the weights might be suboptimal. We tried to dampen the influence of the strong weights at T+1 at later lead times but the improvements to the reliability were minimal.

b) Annual global mean temperature

Constraining by T+1-9, T+2-9, ..., T+9 (smoothed weights by max)

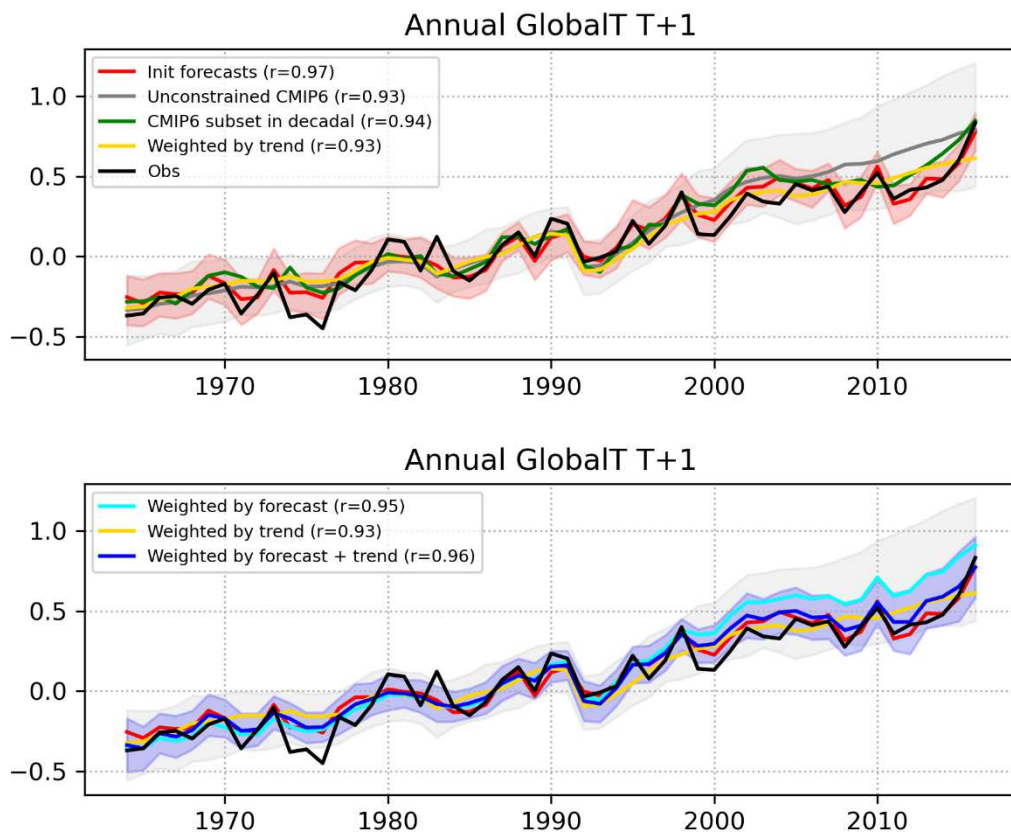


Fig.7 Same as Fig. 4 for annual global mean temperature

The second case study is on annual global mean temperature, where the skill as measured by the ACC is high across all lead times (Fig. 7) because the observed changes are dominated by strong trend. The hindcasts are better at capturing the variability around the trend, notably the temperature anomalies after 1998 which are cooler than the observed long-term trend and not

simulated by the uninitialised CMIP6 realisations. Fig. 5 shows that the trend weighting does not boost the skill (presumably because correlation measures linearity with trend and does not penalise errors in the size of the trend) and that the benefits of the initialisation weighting are seen for T+1 only. This can be seen most clearly in the T+1 projections with the combined weighting which better represents the magnitude of the trend and also the post-1998 cooling. The hindcasts for annual global mean temperature have good signal and more consistency across realisations, and the trend-weighted projections are in better agreement with the observations. These will be important factors in why the hindcasts and weighted projections also have better skill and great reliability than the CMIP6 projections across all lead times.

c) December-February ENSO

Constraining by T+1-9, T+2-9, ..., T+9 (smoothed weights by max)

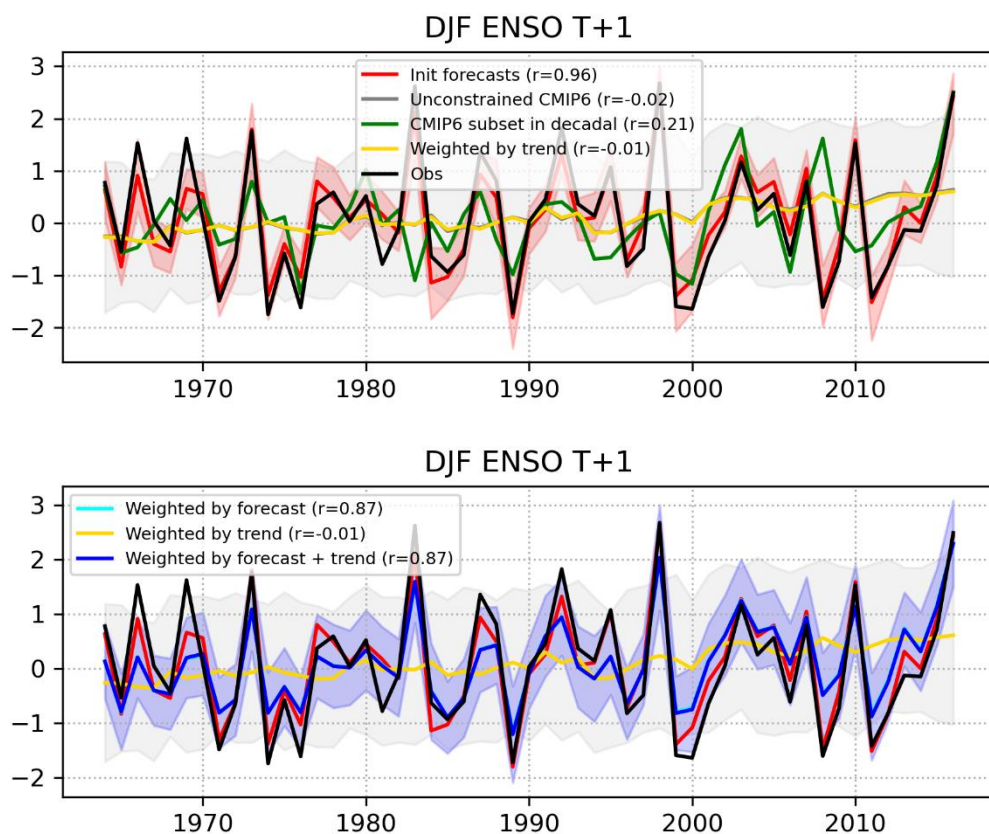


Fig.8 Same as Fig. 4 for DJF ENSO

The third case study is ENSO which is dominated by interannual variability. It also has very little trend and so the effect of trend-weighting is negligible. The hindcasts have very high skill at T+1 (Fig. 8) but this drops off fast so that there is no added value from initialisation by T+4 (see Fig. 1).

470 The hindcasts out-perform the initialisation weighting though this still does very well to realise a correlation of 0.87 for T+1. The hindcasts also have very good reliability whereas the CMIP6 projections are under-dispersive, which is because most CMIP6 models have too little ENSO variability (see Supp Fig. S2). The initialisation weighting reduces the dispersion of an under-dispersive CMIP6 ensemble, so it becomes under-dispersive.

475

d) January-March winter North Atlantic Oscillation Index

Constraining by T+1-9, T+2-9, ..., T+9 (smoothed weights by max)

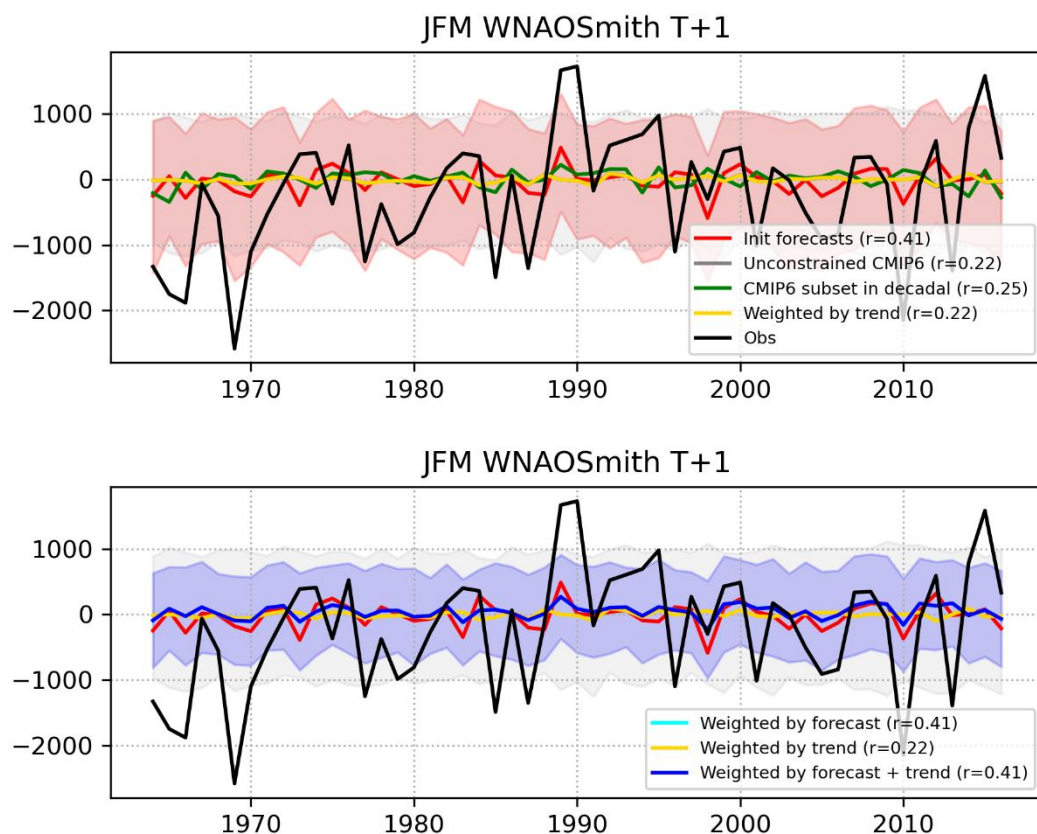


Fig.9 Same as Fig. 4 for January-March NAO

480 The final case study is the winter NAO which has modest skill for T+1 only (see Fig. 9), but is recognised to have signal-to-noise problems. This can be seen in that the observed variations

about the trend are larger than implied by the CMIP6 spread (also see Supp Fig. 2). So the unweighted projections and hindcasts are already under-dispersive before the weighting constrains them further. However, the hindcasts have a modest skill of 0.41 and this is captured
485 by the initialisation weighting. Interestingly, the hindcast skill drops off to zero by T+5 but the combined weighting maintains a small amount of skill out to longer lead times.

The gain in skill at early lead times has worked well. The reliability has worked less well though this is less to do with the weighting and more to do with issues with the climate models such as biases in the mean response and the variability. Overall, the four case studies based on four
490 different variables with four different sets of properties in terms of their sensitivity to manmade forcing, predictability, signal-to-noise problems, and unforced variability have shown a wide range of ways in which the initialisation weighting works. Different projections and forecast products work better in different situations. This will be discussed more in section 6.

495 5. Discussion

Our aim has been to enable users of projections to benefit from the added value of initialised forecasts. The weighting approach has been based on a large-scale index which needs to be skilfully predicted by the decadal forecast system but also chosen to be a key driver of the downstream regional hazard of interest the decision maker. A next step would be to apply the
500 weights determined for the large-scale index to the regional hazard.

Whilst the merging approach has achieved its goals, its usefulness and usability for Climate Services is limited when there is low skill, and when the signal, the variability and the predictable component of internal variability are poorly represented. If the product were to be made operational, several issues would need to be addressed:

- 505 • Variability: Merging climate projections and forecasts involves different climate models with varying quality in their representation of variability. At early lead times, the merged product often fails to match climate projections when internal variability is over- or under-represented. Bias corrections are possible, but care must be taken not to over-constrain variability to the central observational estimate, which itself has considerable
510 spread.
- Forecast signal: skill varies across the different climate models used in the forecasts. Pooling all decadal forecasts and taking the ensemble mean can reduce quality and be biased towards models contributing more realisations.

- Uninitialised forced signal: signal to noise problems can only be resolved by model development though methods to adjust the signal in both the forecast and projections for each contributing model would be helpful in the short term. Ideally, models would contribute a large number of realisations to provide a better estimate of the forced signal from the projections.

The main limitations have been issues inherent to the climate models themselves and need to be addressed before we consider improvements to the merging method. Our approach downweights climate models with variability appreciably higher than that observed, which can be beneficial, but might cause problems for people who have already relied on those projections. In the long term, further model development might reduce these issues but that has proven to be a slow process. New modelling initiatives e.g. Schneider et al. (2024) might expedite the process. The alternative is to correct the biases in the signal and variability in both forecasts and projections before merging. However, this presents a dilemma for methods like ours, which aim to enable users to base their analysis and decisions on a single mainstream data set. Adjusting variability may be relatively straightforward, but adjusting the signals is much more challenging. If bias adjustment is adopted, it is essential that both datasets are corrected consistently prior to merging.

Another study proposed generating probabilistic projections from decadal forecasts and using these to perform Bayesian updating of a prior based on probabilistic long-term projections. The same considerations regarding independence of the weighting scheme and addressing fundamental model errors apply to these emulator-based approaches. In fact, emulator approaches may be a more natural setting for incorporating bias adjustments to both variability and signal.

The current approach is univariate, but it could be extended to a multivariate likelihood function. However, this is likely to expose structural errors which would need to be accounted for. Such an extension could provide the likelihood-weighted equivalent of Donat et al's shadowing approach where they match projections to a combination of recently observed and forecast SSTs patterns after removing the regional mean (similar to our removal of the forced signal).

While the product is not yet operationally ready, it can still be used with caution in certain circumstances. Where skill and reliability are satisfactory, the merged product can be used to indicate the number of years where initialised forecasts would add value to the long-term projections. If the planner's time frame is beyond this time scale focus is resilience, they should

use the weighted projections. For the near term, as skill is similar between the forecast and merged product, the choice will depend on the reliability of the two systems. If one is under-dispersive, then the planner should use the more reliable one. If one is reliable and the other over-dispersive, then the choice depends on the risk tolerance of the planner; a risk averse planner might opt for the over-dispersive option for greater caution. Having said that, data availability is another key consideration. Currently, the decadal forecasts and corresponding hindcast sets are not easy to obtain for the models that contributed after CMIP6. In this case, the planner can decide if the merged product or projections best suit their needs, or if neither is usable.

6. Summary

We have developed a method to create a product based on long-term projections that incorporates the main source of added value of initialised forecasts. This builds on previous approaches that select a subset of long-term projections according to how well they “shadow” the ensemble mean forecast for a large-scale index. Our merged product offers similar benefits to other shadowing methods, such as enabling users to work with one climate data set, the long-term projections, which is a cleaner, less complex data set that does not suffer from drift or initialisation shocks. As the constraints vary with start date, the users benefit from stronger constraints during ‘windows of opportunity’. Additionally, our approach introduces a likelihood-based approach that complements those used in probabilistic climate projections. The method has three advantages over the previous shadowing methods:

- It is applied separately for each lead time, ensuring that information from the forecasts does not over-constrain the projections at lead times beyond the range of predictability. A second step in determining the weights aims to mitigate any resulting loss in consistency across lead times by ensuring once a projection is given a weight it is maintained at that level.
- The likelihood weighting determines an effective ensemble size, eliminating the need of earlier methods to subjectively choose the number of projections to retain.
- The weights are independent of historic trend weights and so can be used in projections that already include such weighting. Here we used a very simple trend weighting for each large-scale index.
- The method is independent of the forced signal and so can be applied to emission scenarios other than the one used in the initialised experiments.

580 For global warming, ENSO, and the predictable subpolar North Atlantic region, the merged product demonstrates skill commensurate with the hindcasts in the first 2, 3, and 5 years. This information on lead times is valuable, as it helps to identify the duration of the near-term period and the onset of the long-term. The reliability is good for annual global mean temperature but under-dispersive for DJF ENSO and JJA SPNA. However, this may be due more to weak signals
585 and our simple univariate trend weighting which could constraint too strongly. A more sophisticated weighting scheme that accounts for structural error would improve the underlying reliability and our method often preserves that level of reliability at longer time scales.

For DJF NAO, which is affected by signal to noise problems, none of the three products are ideal. The hindcasts and merged product exhibit only modest skill at T+1, not much greater than
590 that of the long-term projections. All three products are also under-dispersive. The merged product inherits limitations in either of its parent data set.

The assessment of the three products is very important information for the user to decide which product is most suitable for their purposes. This will also depend on the time scale of the decision, the level of risk the decision maker is prepared to take, but also on practical
595 considerations like data availability.

References

- 600 Befort DJ, O'Reilly CH, Weisheimer A (2020) Constraining Projections Using Decadal
Predictions. *Geophysical Research Letters* 47:e2020GL087900.
<https://doi.org/10.1029/2020GL087900>
- Bethke I, Wang Y, Counillon F, et al (2021) NorCPM1 and its contribution to CMIP6 DCP. *Geoscientific Model Development* 14:7073–7116. <https://doi.org/10.5194/gmd-14-7073-2021>
- 605 Bilbao R, Wild S, Ortega P, et al (2021) Assessment of a full-field initialized decadal climate
prediction system with the CMIP6 version of EC-Earth. *Earth System Dynamics* 12:173–
196. <https://doi.org/10.5194/esd-12-173-2021>
- Boer GJ, Smith DM, Cassou C, et al (2016) The Decadal Climate Prediction Project (DCPP)
contribution to CMIP6. *Geoscientific Model Development* 9:3751–3777.
<https://doi.org/10.5194/gmd-9-3751-2016>
- 610 Boucher O, Servonnat J, Albright AL, et al (2020) Presentation and Evaluation of the IPSL-CM6A-
LR Climate Model. *Journal of Advances in Modeling Earth Systems* 12:e2019MS002010.
<https://doi.org/10.1029/2019MS002010>
- Brunner L, Pendergrass AG, Lehner F, et al (2020) Reduced global warming from CMIP6
projections when weighting models by performance and independence. *Earth System*
615 *Dynamics* 11:995–1012. <https://doi.org/10.5194/esd-11-995-2020>
- Climate Financial Risk Forum (CFRF) (2024) Mobilising adaptation finance to build resilience
- Collier MA, O’Kane TJ, Kitsios V, Sandery PA (2022) CSIRO CAFE-60 submissions to the World
Meteorological Organization operational decadal forecasts and the international multi-
model data exchange. *JSHES* 72:52–57. <https://doi.org/10.1071/ES21024>
- 620 Delgado-Torres C, Donat MG, Pérez-Zanón N, et al (2025) Seamless climate information for the
next months to multiple years: merging of seasonal and decadal predictions, and their
comparison to multi-annual predictions. *EGU sphere* 1–23.
<https://doi.org/10.5194/egusphere-2025-3674>
- Doblas-Reyes FJ, Andreu-Burillo I, Chikamoto Y, et al (2013) Initialized near-term regional
climate change prediction. *Nat Commun* 4:1715. <https://doi.org/10.1038/ncomms2704>
- 625 Eyring V, Bony S, Meehl GA, et al (2016) 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>
- Goldstein M, Rougier J (2004) Probabilistic formulations for transferring inferences from
mathematical models to physical systems. *SIAM JOURNAL ON SCIENTIFIC*
630 *COMPUTING* 26:467–487. <https://doi.org/10.1137/S106482750342670X>
- Harris GR, Sexton DMH, Booth BBB, et al (2013) Probabilistic projections of transient climate
change. *CLIMATE DYNAMICS* 40:2937–2972. <https://doi.org/10.1007/s00382-012-1647-y>

- 635 Hermanson L, Dunstone N, Eade R, Smith D (2024) An ensemble reconstruction of ocean temperature, salinity, and the Atlantic Meridional Overturning Circulation 1960–2021. *Quarterly Journal of the Royal Meteorological Society* 150:98–111. <https://doi.org/10.1002/qj.4587>
- Hersbach H, Bell B, Berrisford P, et al (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* 146:1999–2049. <https://doi.org/10.1002/qj.3803>
- 640 Hu S, Wu B, Wang Y, et al (2023) CAS FGOALS-f3-L Model Datasets for CMIP6 DCP Experiment. *Adv Atmos Sci* 40:1911–1922. <https://doi.org/10.1007/s00376-023-2122-x>
- 645 Intergovernmental Panel on Climate Change (IPCC) (ed) (2023) Decision-Making Options for Managing Risk. In: *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, pp 2539–2654
- Intergovernmental Panel on Climate Change (IPCC) (ed) (2014) Near-term Climate Change: Projections and Predictability. In: *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, pp 953–1028
- 650 Kataoka T, Tatebe H, Koyama H, et al (2020) Seasonal to Decadal Predictions With MIROC6: Description and Basic Evaluation. *Journal of Advances in Modeling Earth Systems* 12:e2019MS002035. <https://doi.org/10.1029/2019MS002035>
- 655 Kharin VV, Boer GJ, Merryfield WJ, et al (2012) Statistical adjustment of decadal predictions in a changing climate. *Geophysical Research Letters* 39:. <https://doi.org/10.1029/2012GL052647>
- Mittal N, Fung F, McSweeney CF, Lowe JA (2025) Identifying user requirements to inform national climate projections in the UK. *Climate Services* 40:100613. <https://doi.org/10.1016/j.cliser.2025.100613>
- 660 Murphy JM, Harris GR, Clark RT (2025) Comparing near-term information from national climate scenarios and initialised decadal predictions. *Clim Dyn* 63:336. <https://doi.org/10.1007/s00382-025-07807-4>
- Nicoli D, Bellucci A, Ruggieri P, et al (2023) The Euro-Mediterranean Center on Climate Change (CMCC) decadal prediction system. *Geoscientific Model Development* 16:179–197. <https://doi.org/10.5194/gmd-16-179-2023>
- 665 O’Kane TJ, Sandery PA, Kitsios V, et al (2021a) CAFE60v1: A 60-Year Large Ensemble Climate Reanalysis. Part I: System Design, Model Configuration, and Data Assimilation. *Journal of Climate* 34:5153–5169. <https://doi.org/10.1175/JCLI-D-20-0974.1>
- O’Kane TJ, Sandery PA, Kitsios V, et al (2021b) CAFE60v1: A 60-Year Large Ensemble Climate Reanalysis. Part II: Evaluation. *Journal of Climate* 34:5171–5194. <https://doi.org/10.1175/JCLI-D-20-0518.1>
- 670 Patrizio CR, Athanasiadis PJ, Smith DM, Nicoli D (2025) Ocean-atmosphere feedbacks key to NAO decadal predictability. *npj Clim Atmos Sci* 8:146. <https://doi.org/10.1038/s41612-025-01027-7>

- 675 Pohlmann H, Müller WA, Bittner M, et al (2019) Realistic Quasi-Biennial Oscillation Variability in Historical and Decadal Hindcast Simulations Using CMIP6 Forcing. *Geophysical Research Letters* 46:14118–14125. <https://doi.org/10.1029/2019GL084878>
- Rougier J, Sexton DMH (2007) Inference in ensemble experiments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365:. <https://doi.org/10.1098/rsta.2007.2071>
- 680 Schneider T, Leung LR, Wills RCJ (2024) Opinion: Optimizing climate models with process knowledge, resolution, and artificial intelligence. *Atmospheric Chemistry and Physics* 24:7041–7062. <https://doi.org/10.5194/acp-24-7041-2024>
- 685 Sexton DMH, Murphy JM, Collins M, Webb MJ (2012) Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. *CLIMATE DYNAMICS* 38:2513–2542. <https://doi.org/10.1007/s00382-011-1208-9>
- Smith DM, Dunstone NJ, Eade R, et al (2025) Mitigation needed to avoid unprecedented multi-decadal North Atlantic Oscillation magnitude. *Nat Clim Chang* 15:403–410. <https://doi.org/10.1038/s41558-025-02277-2>
- 690 Smith DM, Eade R, Scaife AA, et al (2019) Robust skill of decadal climate predictions. *npj Clim Atmos Sci* 2:1–10. <https://doi.org/10.1038/s41612-019-0071-y>
- Smith DM, Scaife AA, Eade R, et al (2020) North Atlantic climate far more predictable than models imply. *Nature* 583:796–800. <https://doi.org/10.1038/s41586-020-2525-0>
- 695 Sospedra-Alfonso R, Merryfield WJ, Boer GJ, et al (2021) Decadal climate predictions with the Canadian Earth System Model version 5 (CanESM5). *Geoscientific Model Development* 14:6863–6891. <https://doi.org/10.5194/gmd-14-6863-2021>
- Thompson V, Dunstone NJ, Scaife AA, et al (2017) High risk of unprecedented UK rainfall in the current climate. *Nat Commun* 8:107. <https://doi.org/10.1038/s41467-017-00275-3>
- 700 Trenberth KE (1997) The Definition of El Niño. *Bulletin of the American Meteorological Society* 78:2771–2778. [https://doi.org/10.1175/1520-0477\(1997\)078%3C2771:TDOENO%3E2.0.CO;2](https://doi.org/10.1175/1520-0477(1997)078%3C2771:TDOENO%3E2.0.CO;2)
- Williams KD, Copsey D, Blockley EW, et al (2018) The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations. *Journal of Advances in Modeling Earth Systems*. <https://doi.org/10.1002/2017MS001115>
- 705 Yukimoto S, Kawai H, Koshiro T, et al (2019) The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component. *Journal of the Meteorological Society of Japan Ser II* 97:931–965. <https://doi.org/10.2151/jmsj.2019-051>

710 **Statements and Declarations**

Funding

This project has received funding from the European Union's Horizon Europe research and 367 innovation programme under grant agreement No 101081460. This work was also funded by 715 U.K. Research and Innovation under the U.K. government's Horizon Europe funding Guarantee (Grant 10060007).

Competing interests

The authors declare that they have no conflict of interest.

720

Author contribution

David Sexton and Carol McSweeney conceived of the idea and chose to base the method on Befort et al. All authors contributed to the interpretation of the results. David Sexton and Jason 725 Lowe developed how the merged product could be user-relevant and how it fitted with current decision-making frameworks. David Sexton wrote the code, made the figures, and wrote the text. Jason Lowe commented on the text.

730 **Data and code availability.**

Code is available at https://github.com/MetOffice/temporal_merge_aspect_paper.

Data is available from DOI 10.5281/zenodo.20066594.

Tables

735 **Table 1.** Details about the models that contribute to the decadal forecasts

Institute	Model	Does model have > 1 CMIP6 realisation?	Number of realisations	Reference
Canadian Centre for Climate Modelling and Analysis (CCCma)	CanESM5	Y	10	Sospedra-Alfonso et al. 2021
Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC)	CMCC-CM2-SR5	N	40	Nicoli et al. 2023
Barcelona Supercomputing Center (BSC)	EC-Earth3.3	N	10	Bilbao et al. 2021
Chinese Academy of Sciences (CAS)	FGOALS-f3-L	N	9	Hu et al. 2023
Commonwealth Scientific and Industrial Research Organisation (CAFE)	GFDL-CM2	N	10	O’Kane et al. 2021a, b; Collier et al. 2022
Met Office Hadley Centre (MOHC)	HadGEM3-GC31-MM (DePreSys4)	Y	10	Williams et al. 2018; Hermanson et al. 2024
Institut Pierre-Simon Laplace (IPSL)	IPSL-CM6A-LR	N	10	Boucher et al. 2020
MIROC consortium (MIROC)	MIROC6	Y	10	Kataoka et al. 2020
Max Planck Institute for Meteorology (MPI-M)	MPI-ESM-HR	N	10	Pohlmann et al. 2019
Meteorological Research Institute (MRI)	MRI-EMS2-0	Y	10	Yukimoto et al. 2019
NorESM climate modelling consortium (NCC)	NorCPM1_i1	N	10	Bethke et al. 2021
NorESM climate modelling consortium (NCC)	NorCPM1_i2	N	10	Bethke et al. 2021

Table 2. List of CMIP6 models used and the institutions that provided simulations from those models. Information regarding the available scenario experiments for the CMIP6 models is available at

https://pcmdi.llnl.gov/CMIP6/ArchiveStatistics/esgf_data_holdings/ScenarioMIP/index.html

Model and Realisation	Institution
ACCESS-CM2 i1p1f1	Commonwealth Scientific and Industrial Research Organization (Australia)
ACCESS-ESM1-5 i1p1f1	
AWI-CM-1-1-MR i1p1	Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (Germany)
AWI-ESM-1-REcoM i1p1	
BCC-CSM2-MR i1p1f1	Beijing Climate Center (China)
CAMS-CSM1-0 i1p1f1	Chinese Academy of Meteorological Sciences (China)
CAS-ESM2-0 i1p1	Chinese Academy of Sciences (China)
FGOALS-g3 i1p1f1	
CESM2 i1p1f1	National Center for Atmospheric Research (USA)
CESM2-FV2 i2p2	
CESM2-WACCM i1p1f1	
CMCC-CM2-SR5 i1p1	Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC)
CMCC-ESM2 i1p1	
CNRM-CM6-1 i1p1f2	Centre National de Recherches Meteorologiques, Centre Europeen de Recherche et de Formation Avancee en Calcul Scientifique (France)
CNRM-CM6-1-HR i1p1f2	
CNRM-ESM2-1 i1p1f2	
CanESM5 i1p1f1, i1p2f1	Canadian Centre for Climate Modelling and Analysis (Canada)
CanESM5-1 i1p1f1, i1p2f1	
CanESM5-CanOE i1p2f1	
FIO-ESM-2-0 i1p1f1	First Institute of Oceanography, Qingdao National Laboratory for Marine Science and Technology (China)
GFDL-CM4 i1p1f1	Geophysical Fluid Dynamics Laboratory, National Oceanic and Atmospheric Administration (USA)
GFDL-ESM4 i1p1f1	
GISS-E2-1-G i1p1f2, i1p3f1, i1p5f1	Goddard Institute for Space Studies, National Aeronautics and Space Administration (USA)
GISS-E2-1-H i1p1f2, i1p3f1	
GISS-E2-2-G i1p3f1	
HadGEM3-GC31-LL i1p1f3	Met Office Hadley Centre (UK)
HadGEM3-GC31-MM i1p1f3	
UKESM1-0-LL i1p1f2	
IITM-ESM i1p1f1	Centre for Climate Change Research, Indian Institute of Tropical Meteorology (India)
INM-CM4-8 i1p1f1	Institute for Numerical Mathematics, Russia Academy of Science (Russia)
INM-CM5-0 i1p1f1	Institute for Numerical Mathematics, Russia Academy of Science (Russia)
IPSL-CM6A-LR i1p1f1	Institut Pierre Simon Laplace, Paris 75252, France
KACE-1-0-G i1p1f1	National Institute of Meteorological Sciences/Korea Meteorological Administration (Korea)
MIROC-ES2H i1p4f2	

MIROC-ES2L i1p1f2	Japan Agency for Marine-Earth Science and Technology (Japan)
MIROC6 i1p1f1	
MPI-ESM1-2-HR i1p1f1	Max Planck Institute for Meteorology (Germany)
MRI-ESM2-0 i1p1f1, i2p1f1	Meteorological Research Institute, Japan Meteorological Agency (Japan)
NESM3 i1p1f1	Nanjing University of Information Science and Technology (China)
NorESM2-LM i1p1f1	Norwegian Climate Center (Norway)
NorESM2-MM i1p1f1	
TaiESM1 i1p1f1	Research Center for Environmental Changes (Taiwan)

Table 3. List of indices of large-scale drivers and their definitions

Short Name	Long Name	Definition	Reference
JJA SPNA	Summer subpolar North Atlantic temperature	Average of sea surface temperature over 60-10°W, 50-60°N.	Smith et al. 2019
Annual Global T	Annual global mean temperature	Global average of surface air temperature.	N/A
DJF ENSO	Dec-Feb ENSO	Nino 3.4 index. Average of sea surface temperature over 170-120°W, 5°S-5°N.	Trenberth 1997
JFM Winter NAO	Late winter NAO	Difference of MSLP: Azores region 28-20°W, 36-40°N minus Iceland region 25-16°W, 63-70°N	Smith et al. 2020

Captions

Fig. 1 Residual correlation for decadal forecasts of JJA SPNA. The correlation is estimated using the approach in section 3. This is an adaptation of the original method in Smith et al (2019), with a different way to remove the forced signal

Fig.2 The individual realisations in the CMIP6 projections (thin lines) compared to the reanalysis time series (light grey) and the average over the pooled set (dark blue) for each large-scale index (each row). The left panels show unfiltered time series. In the right panels, only the low-pass filtered averages are shown for each ensemble. The magenta lines show ensembles that are not consistent with the reanalysis trend, otherwise the lines are light blue. The red line shows the trend line estimated for the reanalysis time series from 1991-2020, extrapolated back to 1960

Fig.3 Estimation of the ensemble mean residual signal for each decadal forecast for summer SPNA. (left panels) The drift-corrected anomaly time series for each realisation of the forecast (grey lines) and their ensemble average (blue line) compared to the smooth estimated forced signal from the CMIP6 projections appropriate for that forecast model (red line) and the reanalysis (black line). (right panels) Comparison of the ensemble mean residuals with the reanalysis residuals after the uninitialised short- and long-term signals are regressed out

Fig.4 For summer subpolar North Atlantic temperatures over 1963-2016, a comparison of the merged product (blue line and plume) with the original CMIP6 data (grey line and plume), reanalysis (black line) and the initialised hindcasts (red line and plume). Plumes show the 10th-90th percentile confidence interval. A simple trend weighting is applied to the CMIP6 data beforehand (yellow line) and the effect on the merged product without the trend weighting is also shown (light blue line)

Fig.5 Decision-relevant metrics as a function of lead time for the four large-scale indices (each column). The metrics are anomaly correlation (top row; note that the y-range changes for each index), spread-over-error (middle row), and fraction within the 80% confidence interval (bottom row)

Fig.6 The effective ensemble size as a function of lead time and start year for each of the four large-scale indices

Fig.7 Same as Fig. 4 for annual global mean temperature

Fig.8 Same as Fig. 4 for DJF ENSO

Fig.9 Same as Fig. 4 for January-March NAO

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TemporalMergingSuppInfo.docx](#)