



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE DELLA RICERCA

Alma Mater Studiorum Università di Bologna Archivio istituzionale della ricerca

North Atlantic climate far more predictable than models imply

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

North Atlantic climate far more predictable than models imply / Smith D.M.; Scaife A.A.; Eade R.; Athanasiadis P.; Bellucci A.; Bethke I.; Bilbao R.; Borchert L.F.; Caron L.-P.; Counillon F.; Danabasoglu G.; Delworth T.; Doblas-Reyes F.J.; Dunstone N.J.; Estella-Perez V.; Flavoni S.; Hermanson L.; Keenlyside N.; Kharin V.; Kimoto M.; Merryfield W.J.; Mignot J.; Mochizuki T.; Modali K.; Monerie P.-A.; Muller W.A.; Nicoli D.; Ortega P.; Pankatz K.; Pohlmann H.; Robson J.; Ruggieri P.; Sospedra-Alfonso R.; Swingedouw D.; Wang X.; Wildhijer S.; Yeager S.; Yang X.; Zhang L. - In: NATURE. - ISSN 0028-0836. - ELETTRONICO. - 583:7818(2020), pp. 796-800. [10.1038/s41586-020-2525-0]

This version is available at: <https://hdl.handle.net/11365/769573> since: 2020-08-29

Published:

DOI: <http://doi.org/10.1038/s41586-020-2525-0>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Smith, D.M., Scaife, A.A., Eade, R. et al. *North Atlantic climate far more predictable than models imply*. Nature 583, 796–800 (2020).

The final published version is available online at: <https://doi.org/10.1038/s41586-020-2525-0>

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

Climate models underpredict North Atlantic atmospheric circulation changes

D. M. Smith¹, A. A. Scaife^{1,2}, R. Eade¹, P. Athanasiadis³, A. Bellucci³, I. Bethke^{4,5}, R. Bilbao⁶, L. F. Borchert⁷, L.-P. Caron⁶, F. Counillon^{4,5}, G. Danabasoglu⁸, T. Delworth⁹, F. J. Doblas-Reyes^{6,10}, N. J. Dunstone¹, V. Estella-Perez⁷, S. Flavoni⁷, L. Hermanson¹, N. Keenlyside^{4,5}, V. Kharin¹¹, M. Kimoto¹², W. J. Merryfield¹¹, J. Mignot⁷, T. Mochizuki^{13,14}, K. Modali¹⁵, P.-A. Monerie¹⁶, W. A. Müller¹⁵, D. Nicolí³, P. Ortega⁶, K. Pankatz¹⁷, H. Pohlmann^{15,17}, J. Robson¹⁶, P. Ruggieri³, R. Sospedra-Alfonso¹¹, D. Swingedouw¹⁸, Y. Wang⁴, S. Wild⁶, S. Yeager⁸, X. Yang⁹ and L. Zhang⁹

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

²*College of Engineering, Mathematics and Physical Sciences, Exeter University, UK*

³*Centro Euro-Mediterraneo sui Cambiamenti Climatici, Bologna, Italy*

⁴*Nansen Environmental and Remote Sensing Center and Bjerknes Centre for Climate Research, Bergen, Norway*

⁵*Geophysical Institute, University of Bergen and Bjerknes Centre for Climate Research, Bergen, Norway*

⁶*Barcelona Supercomputing Center, Jordi Girona 29 - 08034 Barcelona, Spain*

⁷*Sorbonne Universités, LOCEAN Laboratory, Institut Pierre Simon Laplace (IPSL), Paris, France*

⁸*National Center for Atmospheric Research, Boulder, CO, USA*

⁹*Geophysical Fluid Dynamics Laboratory, Princeton University, Princeton, NJ, USA*

¹⁰*ICREA, Barcelona, Spain*

¹¹*Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada,*

22 *Victoria, British Columbia, Canada*

23 ¹²*Atmosphere and Ocean Research Institute, University of Tokyo, Kashiwa, Japan*

24 ¹³*Department of Earth and Planetary Sciences, Kyushu University, Fukuoka, Japan*

25 ¹⁴*Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan*

26 ¹⁵*Max-Planck-Institut für Meteorologie, Bundesstraße 53, 20146 Hamburg, Germany*

27 ¹⁶*National Centre for Atmospheric Science, Department of Meteorology, University of Reading,*
28 *Reading RG6 6BB, UK*

29 ¹⁷*Deutscher Wetterdienst, Bernhard-Nocht-Str. 76, Hamburg, Germany*

30 ¹⁸*CNRS-EPOC, Université de Bordeaux, Pessac, France*

31 *Corresponding author: Doug Smith, doug.smith@metoffice.gov.uk*

32 **Abstract**

33 **Quantifying signals and uncertainties in climate models is essential for climate change de-**
34 **tection, attribution, prediction and projection¹⁻³. Although inter-model agreement is high**
35 **for large-scale temperature signals, dynamical changes in atmospheric circulation are very**
36 **uncertain⁴, leading to low confidence in regional projections especially for precipitation over**
37 **the coming decades^{5,6}. Furthermore, model simulations with tiny differences in initial condi-**
38 **tions suggest that uncertainties may be largely irreducible due to the chaotic nature of the cli-**
39 **mate system⁷⁻⁹. However, climate projections are difficult to verify until further observations**
40 **become available. Here we assess retrospective climate predictions of the last six decades**

41 **and show that decadal variations in north Atlantic winter atmospheric circulation are highly**
42 **predictable. Crucially, climate models underestimate the predictable signal by an order of**
43 **magnitude and skill is achieved despite a lack of agreement between individual model simu-**
44 **lations. Consequently, skilful climate predictions of European and eastern North American**
45 **winters are possible but require 100 times more ensemble members than would perfect mod-**
46 **els and post-processing to overcome underestimated teleconnections. Our results highlight**
47 **the pressing need to understand why the signal-to-noise ratio is too small in climate models¹⁰,**
48 **and the extent to which correcting this model error would reduce uncertainties in regional**
49 **climate change on timescales beyond a decade.**

50 Global climate models are used extensively to understand the drivers of past climate variabil-
51 ity and change, and to predict what is likely to happen in the future¹⁻³. Underpinning this is a need
52 for accurate estimates of signals and associated uncertainties in climate model simulations in order
53 to distinguish between different causes of past climate change, and to provide reliable confidence
54 limits on future projections. Uncertainties are typically partitioned into three sources¹¹: scenario
55 uncertainty arising from an imperfect knowledge of external forcing factors, including changes
56 in greenhouse gases, ozone, anthropogenic and volcanic aerosols, and solar irradiance; modelling
57 uncertainty arising from the fact that different models respond differently to the same radiative
58 forcing; and internal variability of climate that would occur in the absence of any external forcing.

59 Climate projections for many regions are currently highly uncertain, especially for atmo-
60 spheric circulation^{4,12} and related impacts, including precipitation^{5,6}. This is particularly well

61 illustrated by the fact that modelling^{13,14} and internal variability^{7,8} uncertainties are each large
62 enough to allow opposite projections of European winters, especially for the coming decades.
63 Whilst modelling uncertainties might be reduced as models improve, internal variability uncer-
64 tainties have been interpreted to be largely irreducible⁷⁻⁹ suggesting that confident projections of
65 European winters may never be possible. However, such conclusions assume that signals and un-
66 certainties diagnosed from climate models are correct. Although multi-decadal and longer climate
67 projections are difficult to verify until future observations become available, signals over the first
68 10 years can be more robustly evaluated using retrospective decadal predictions (hereafter referred
69 to as hindcasts).

70 We use a very large multi-model ensemble of decadal hindcasts from the Coupled Model
71 Intercomparison Project (CMIP) phases 5¹⁵ and 6¹⁶. We focus on the boreal winter period (De-
72 cember to March) averaged over forecast years 2 to 9 to avoid seasonal to annual predictability
73 and focus on decadal timescales. We use hindcasts starting each year over the period 1960 to 2005
74 from 6 CMIP5 and 8 CMIP6 modelling systems, giving a total of 169 ensemble members which
75 are weighted equally (see Methods, Table1). Hence our total hindcast dataset comprises 77,740
76 (46 start dates times 169 ensemble members times 10 years) years of model integrations to provide
77 robust statistics.

78 To compare with uncertainties in climate projections^{5,7,8,13,14} we focus on European winters
79 which are largely controlled by the North Atlantic Oscillation (NAO), the leading mode of atmo-
80 spheric circulation variability in the north Atlantic¹⁷. The NAO represents the meridional gradient

81 in mean sea level pressure (mslp), typically measured as the difference in pressure between the
82 Azores and Iceland. Its positive (negative) phase reflects an increased (reduced) pressure gradi-
83 ent driving stronger (weaker) mid-latitude westerly winds with increased (reduced) storminess,
84 and a northward (southward) shift of the jet stream. Impacts of the NAO are characterised by a
85 quadrupole pattern, with a positive (negative) NAO driving warmer, wetter (colder, drier) condi-
86 tions in northern Europe and south-east North America along with colder, drier (warmer, wetter)
87 conditions in southern Europe and north-east North America.

88 We assess skill using two different measures (see Methods): anomaly correlation coefficient
89 (ACC) which measures the phase of variability, and mean-squared-skill-score (MSSS) which mea-
90 sures the amplitude of variability. We find significant skill for decadal predictions of winter mslp in
91 most regions, including the north Atlantic, when measured by the ACC between the 169-member
92 ensemble mean and observations (Figure 1a). However, skill is much lower especially in the north
93 Atlantic when measured by the MSSS or the ACC of a smaller (10-member, typical of individual
94 prediction systems¹⁶) ensemble mean (Figure 1 b and c). Timeseries from the observations and
95 each model ensemble member consist of a predictable component (the signal) and unpredictable
96 internal variability (the noise). The discrepancy in skill between ACC and MSSS, and the need for
97 a large ensemble, arise because the signal-to-noise ratio is too small in the models compared to
98 observations^{10, 18, 19}. Hence, skill is low in a 10-member ensemble mean because a larger ensemble
99 is required to reduce the noise and extract the predicted signal. In contrast, the signal resulting from
100 a large ensemble mean may capture the correct phase of observed variability giving a significant
101 ACC, but its amplitude will be much too small resulting in a low MSSS.

102 Errors in the signal-to-noise ratio can be quantified by comparing the predictable compo-
103 nents (the predictable fraction of the total variability) in observations and models. The ratio of
104 predictable components^{10,18,20} (RPC, see Methods) is expected to be one for a perfect forecasting
105 system; values greater than one show where the signal-to-noise ratio is erroneously too small in
106 models. Consistent with differences in ACC and MSSS we find RPC is greater than one almost
107 everywhere where there is skill in ACC, and especially in the north Atlantic (Figure 1d).

108 The NAO exhibits marked decadal variability²¹ with a strong increase from the 1960s to the
109 1990s and a decrease thereafter (Figure 2a, black curve). The raw ensemble mean forecast shows
110 virtually no signal (Figure 2a, red curve), and the observations generally lie within the model
111 uncertainties (shading showing the 5-95% range diagnosed from the ensemble spread), although
112 the extreme values in the early 1960s and late 1980s are not well-captured by models in agreement
113 with other studies^{22,23}. Taken at face value, as is done for climate projections^{5,7,8,14}, the small
114 model signal and much larger spread would imply little ability to predict the NAO and a large
115 component of unpredictable internal variability. However, by comparing with observations we find
116 significant correlation skill of the ensemble mean (ACC=0.48, p=0.02), while persistence provides
117 a poor forecast (ACC=0.1). Hence, skilful climate model predictions of the NAO are possible using
118 the ensemble mean, but the signal-to-noise ratio is too small (RPC=4.2) and its variance must be
119 calibrated to provide realistic forecasts¹⁹.

120 Rescaling the ensemble mean time-series to have the same variance as the observations re-
121 veals that the predictions do capture the observed increase from the 1960s to 1990s and decrease

122 thereafter (Figure 2b). However, even with 169 ensemble members (Figure 2b thin red curve)
123 there are large interannual variations that are not expected or observed in 8-year rolling means. We
124 therefore create a larger lagged ensemble by taking the average of the four latest forecasts avail-
125 able at each start date (giving 676 members, Figure 2b thick red curve, see Methods). This reveals
126 that the NAO is highly predictable on decadal timescales ($ACC=0.79$, $p<0.01$) in stark contrast to
127 the lack of predictability implied by the standard interpretation of raw model output (Figure 2a).
128 Importantly, the signal-to-noise ratio is much too small in the models ($RPC=11$, $p=0.02$). The
129 total 8-year variability of the NAO in individual model members (standard deviation = 1.7 to 2.6
130 hPa, 5-95% range, year 2-9 hindcasts) is not significantly different to the observations (2.4hPa).
131 Hence the predictable signal (see Methods) is underestimated by an order of magnitude in the
132 model ensemble. Since the standard error of the ensemble mean is reduced by the square root of
133 the ensemble size, the ensemble required to extract the signal is 100 times larger than it would be
134 for perfect models.

135 The fact that the NAO signal is much too weak in models implies that the impacts of the
136 NAO will be underestimated relative to other factors such as greenhouse gases. Hence in regions
137 influenced by the NAO the ensemble mean will not reflect the true balance of driving factors and
138 simply inflating its variance to be the same as observed will not correct the error. A potential so-
139 lution is to post-process the model output by selecting a subset of (20) ensemble members from
140 the lagged ensemble (of 676 members) whose simulated NAO is closest in sign and magnitude
141 to the ensemble mean NAO after adjusting this to take into account the underestimated signal.
142 These members contain close-to the correct magnitude of the forecast NAO whilst retaining influ-

143 ences from greenhouse gases and other sources. We refer to this procedure as “NAO-matching”
144 (see Methods) and note that it builds on previous techniques^{24,25} by using the models as much as
145 possible instead of observed relationships which may not be causal or robust.

146 We investigate this technique first for forecasts of Atlantic Multidecadal Variability (AMV,
147 see Methods). AMV is thought to be one of the most predictable aspects of decadal climate²⁶, yet
148 the lagged ensemble mean does not capture the correct timing of the minimum in the late 1980s
149 (Figure 2c). NAO-matching captures the minimum and subsequent rapid warming in much bet-
150 ter agreement with observations (Figure 2d) consistent with evidence that AMV is at least partly
151 forced by the NAO^{27–29}. We find similar improvements for northern European rainfall: the lagged
152 ensemble mean is not significantly skilful and the observations lie outside the modelled uncer-
153 tainties in the 1960s and 1980s (Figure 2e), whereas the NAO-matched ensemble is significantly
154 skilful ($ACC=0.72$, $p<0.01$) and captures the observed increase from the 1960s to late 1980s and
155 decrease thereafter. As expected, these improvements are not seen by simply adjusting the variance
156 of the ensemble mean (Supplementary Figure S1).

157 Forecasts of extreme decades would be of particular value since they could enable action
158 to be taken in advance to avoid the most severe climate impacts³⁰. We therefore investigate the
159 extreme positive NAO period between 1986 and 1997 (8-year means starting 1986 to 1990, Fig-
160 ure 2a). Consistent with the above results, the raw lagged ensemble mean shows virtually no signal
161 compared to observed variability (Figure 3 a, b, c compared to d, e, f). Adjusting its variance to be
162 equal to the observed variance (Figure 3 g, h, i) reveals that the forecasts do capture the positive

163 NAO (as expected from Figure 2b), but the expected impacts are underestimated, especially for
164 temperature and northern European precipitation. However, the NAO-matched forecast (Figure 3
165 j, k, l) shows a clear improvement and captures the expected quadrupole pattern with warm, wet
166 (cold, dry) anomalies in northern Eurasia and south-east North America (northern Africa and parts
167 of southern Europe, and north-east North America), as well as low pressure across the Arctic. Sim-
168 ilar improvements from NAO-matching are found for trends and for skill measured over all of the
169 hindcasts (Supplementary Figures S3-S4).

170 We have shown that the winter NAO and related impacts on Europe and eastern North Amer-
171 ica are highly predictable on decadal timescales. AMV is usually believed to be a major source of
172 decadal prediction skill^{26,31}. However, we find that predictions of AMV can be improved by using
173 the forecast NAO (Figure 2c,d), whereas predictions of the NAO are degraded by selecting the
174 most skilful AMV ensemble members (Supplementary Figure S5). This suggests that the NAO is
175 not solely driven by AMV. Hence other potential influences, including for example the tropics³²⁻³⁴,
176 warrant further investigation.

177 Crucially we find that the NAO signal is underestimated by an order of magnitude in the
178 model ensemble. This adds to an increasing body of evidence that the signal-to-noise ratio is
179 too small in climate models, seen on seasonal^{20,35-37}, interannual³⁸ and decadal^{19,39} timescales.
180 Consequently, the real world is more predictable than climate models suggest^{10,18} and uncer-
181 tainties diagnosed from raw model simulations are too large. The cause of this error is not yet
182 known, though there are several hypotheses including weak teleconnections to the quasi-biennial

183 oscillation⁴⁰, lack of persistence in the NAO^{41,42} and in weather regimes⁴³, unresolved ocean at-
184 mosphere interactions⁴⁴ and weak transient eddy feedback⁴⁵.

185 A key question is whether climate models also underestimate signals on timescales beyond a
186 decade. There is some evidence that the atmospheric circulation response to Arctic sea ice loss⁴⁶,
187 and to external factors¹⁰ including volcanic eruptions, solar variations and ozone changes, are too
188 weak in models. Models also appear to underestimate the magnitude of multi-decadal temperature
189 variability^{47,48} especially for the north Atlantic^{49,50}. Furthermore, model-simulated winter climate
190 change signals in the north Atlantic increase substantially as resolution increases⁵¹, consistent
191 with the suggestion that eddy feedbacks are inadequately resolved⁴⁵. If this is robust, treating
192 current model simulations at face value is giving misleading conclusions about uncertainties and
193 irreducible internal variability.

194

195

196

197 **Methods**

198 **Observations and models.** Near surface temperature observations are computed as the average
199 of HadCRUT4⁵², NASA-GISS⁵³ and NCDC⁵⁴. Precipitation observations are taken from GPCC⁵⁵

200 and mean sea level pressure is taken from HadSLP2⁵⁶.

201 We assess a large multi-model ensemble (169 members, Table 1) of decadal predictions from
202 14 modelling systems using hindcasts starting each year from 1960 to 2005 from the Coupled
203 Model Intercomparison Project (CMIP5) phases 5¹⁵ and 6¹⁶. We found no significant difference
204 in NAO correlation skill between the CMIP5 and CMIP6 ensembles and focus on the combined
205 ensemble to obtain the most robust statistics. We create ensemble means by taking the equally-
206 weighted average of all ensemble members and assess rolling 8-year boreal winter (December to
207 March) means defined by calendar years 2 to 9 from each start date. The forecasting systems
208 start between 1st of November and January each year, giving a lead time of at least a year before
209 the assessed forecast period to focus on decadal timescales and avoid predictability arising from
210 seasonal to annual variability. Both halves of the 8-year period contribute to skill (NAO ACC =
211 0.57 and 0.45, $p=0.03$, for forecast years 2 to 5 and 6 to 9 respectively). Both observations and
212 models were interpolated to a 5° longitude by 5° latitude grid before comparison.

213 **Indices.** The North Atlantic Oscillation (NAO) index is calculated as the difference in mean sea
214 level pressure between two small boxes located around the Azores (28-20°W, 36-40°N) and Iceland
215 (25-16°W, 63-70°N) with the average over the whole time series removed to create anomalies³⁸.
216 Atlantic Multidecadal Variability (AMV) is calculated as the near-surface temperature in the North
217 Atlantic (80-0°W, 0-60°N) minus the global average (60°S-60°N)⁵⁷. European rainfall is averaged
218 over the box 10°W-25°E, 55-70°N. All forecasts indices are based on the ensemble mean.

219 **Forecast quality and uncertainty measures.** Model biases and drifts are treated by computing
 220 anomalies relative to climatology for each model computed over all hindcasts, and comparing with
 221 observed anomalies computed over the same period. Although there are many ways to measure
 222 forecast quality, we focus on those that illustrate the underestimated model signals by using the
 223 following:

$$\text{Pearson anomaly correlation coefficient ACC} = \frac{\sum_{i=1}^N (f_i - \bar{f})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^N (f_i - \bar{f})^2 \sum_{i=1}^N (o_i - \bar{o})^2}} \quad (1)$$

$$\text{Mean-squared-skill-score MSSS} = 1 - \frac{\sum_{i=1}^N (f_i - o_i)^2}{\sum_{i=1}^N (\bar{o} - o_i)^2} \quad (2)$$

$$\text{Ratio of predictable components RPC} = \frac{\sigma_{sig}^o / \sigma_{tot}^o}{\sigma_{sig}^f / \sigma_{tot}^f} = \frac{ACC}{\sigma_{sig}^f / \sigma_{tot}^f} \quad (3)$$

$$\text{Ratio of predictable signals} = \frac{\sigma_{sig}^o}{\sigma_{sig}^f} = RPC \frac{\sigma_{tot}^o}{\sigma_{tot}^f} \quad (4)$$

224 where N is the number of hindcast start dates, f_i and o_i are the ensemble mean forecast and
 225 observations at each time, and the overbar represents the average over all times. σ_{sig} and σ_{tot} are
 226 the expected standard deviations of the predictable signal and total variability, with superscripts o
 227 and f for the observations and forecasts respectively. For the forecasts, σ_{sig} and σ_{tot} are computed
 228 from the ensemble mean and individual members respectively.

229 ACC measures the ability to predict the phase of variability, whereas MSSS measures the
 230 magnitude of errors relative to a climatological forecast. For a perfect forecasting system RPC
 231 should equal one. Note that RPC is not computed where the ACC is negative, and that the above
 232 formula likely gives a lower bound^{10,18}.

233 Uncertainties in raw model forecasts are computed from the ensemble standard deviation

234 for each start date. Uncertainties in variance adjusted and NAO-matched forecasts are computed
235 from the root-mean-square error between the ensemble mean and the observations as required for
236 reliable forecasts⁵⁸.

237 We note that it is theoretically possible for the multi-model RPC to be larger than for individ-
238 ual models if time dependent model biases⁵⁹ or teleconnection errors reduce the model signal more
239 than the correlation with observations. Assessing this thoroughly would require large ensembles
240 of individual model hindcasts which are not available. However, assessing the largest individual
241 model ensemble available (NCAR CESM1.1 with 40 members per year, giving 160 lagged mem-
242 bers, Table 1) does not support this hypothesis: the NCAR RPC of 6.2 is not significantly different
243 from the average RPC of multi-model ensembles of the same size (4.8 averaged over 1000 ran-
244 dom samples, with 5-95% range 1.3 to 7.4). Furthermore, the statistics presented in this study are
245 appropriate for multi-model ensemble forecasts.

246 We further note that there is some evidence that the predictability of the NAO may vary
247 on multi-decadal timescales⁶⁰, though this is not robust across models⁶¹. Our results are statisti-
248 cally significant for the hindcast period available, but longer hindcasts that include more cycles of
249 decadal variability would be beneficial for future studies.

250 **Lagged ensemble.** Consecutive 8-year means contain 7 identical years. Hence large interannual
251 variations, as seen in 169-member ensemble mean NAO forecasts (Figure 2b), are not expected.
252 They occur because the signal to noise ratio is too small in models and consecutive decadal pre-
253 dictions consist of independent model simulations that are dominated by different samples of the

254 noise. Ideally additional ensemble members would be used to reduce the noise further, but these
 255 are not available. Instead we create a lagged ensemble by combining the required forecast with the
 256 previous three i.e. the year 2-9 forecasts starting in 1963 are combined with the year 2-9 forecasts
 257 starting in 1962, 1961 and 1960 giving a total of 676 members (169 members time 4 start dates).
 258 The previous forecasts are sub-optimal because they do not cover exactly the same forecast period,
 259 and rely on the persistence of running 8-year means. Hence there is a trade off between reducing
 260 the noise with additional members and potentially degrading the skill by relying on persistence.
 261 In the current generation of climate models the benefit in reducing the noise far outweighs the
 262 degradation from using persistence. We present results for the combination of 4 lagged forecasts,
 263 but find similar levels of skill for other combinations (NAO ACC = 0.71 and 0.78 for combining 3
 264 and 5 lagged forecasts respectively). A similar technique relying on persistence of the predictor re-
 265 cently proved to strongly reduce the noise in decadal predictions of summer temperature extremes
 266 over land⁶².

267 **NAO-matching.** At any location that is influenced by the NAO we can write

$$O = O_{NAO} + O_{OTHER} + \epsilon^o \quad (5)$$

$$F^k = F_{NAO}^k + F_{OTHER}^k + \epsilon^k \quad (6)$$

$$\hat{F} = \hat{F}_{NAO} + \hat{F}_{OTHER} + \hat{\epsilon} \quad (7)$$

268 where O , F^k and \hat{F} are the observed, forecast ensemble member k and forecast ensemble mean
 269 values of a meteorological variable (e.g. temperature, rainfall, pressure). The subscript *NAO* refers
 270 to the portion that is related to the NAO, the subscript *OTHER* refers to the portion related to
 271 other predictable drivers (including greenhouse gases and sea surface temperatures unrelated to

272 the NAO) and ϵ is an unpredictable residual. Because the predictable NAO signal is too small in
273 models, the mean of a very large ensemble is required for skilful NAO predictions (Figure 2b).
274 However, the magnitude of the ensemble mean NAO is much too small (Figure 2a) and therefore
275 \hat{F}_{NAO} will be severely underestimated.

276 One approach to overcoming model deficiencies uses regressions between model hindcasts
277 and observations^{25, 63–65}, which effectively replaces the erroneous \hat{F}_{NAO} with the observed value
278 O_{NAO} . Whilst this can give very good results, it relies on O_{NAO} estimated from the observations
279 being robust and describing a causal relationship between the NAO and remote regions. This
280 approach is less attractive on decadal than seasonal timescales because O_{NAO} is potentially more
281 affected by sampling errors from the relatively small hindcast period.

282 An alternative approach²⁴ replaces the underestimated \hat{F}_{NAO} with more realistic F_{NAO}^k by
283 selecting from the full ensemble a smaller set of members that have the required magnitude of
284 the NAO. These members contain close-to the correct magnitude of the required NAO and its
285 teleconnections whilst retaining other influences. Hence, \hat{F}_{NAO} for this selected ensemble will be
286 larger than that of the full ensemble, thereby increasing the signal. Because the selected ensemble
287 is smaller the remaining noise will not be reduced as much as in the full ensemble. However,
288 the selection process transfers variability from what would be considered as noise in a random
289 ensemble into \hat{F}_{NAO} , thereby reducing $\hat{\epsilon}$ in the selected ensemble. Hence, in regions affected by
290 the NAO the increase in signal is likely to be larger than the reduced suppression of the remaining
291 noise, thereby increasing the signal to noise ratio and improving the skill.

292 In the previous seasonal forecast study²⁴ the required NAO was obtained based on observed
293 relationships with potential drivers. However, on decadal timescales such relationships are not
294 well-established and are more likely to be affected by sampling errors. We therefore take the re-
295 quired NAO to be the ensemble mean forecast NAO but adjusted to account for the underestimation
296 of the predictable signal. This is achieved by multiplying the ensemble mean NAO by the ratio of
297 predictable signals (equation 4). To avoid overfitting to observations we compute the ratio of pre-
298 dictable signals for each hindcast start date separately using a cross-validation approach in which
299 the required hindcast and those on either side are omitted. Our conclusions are robust to omit-
300 ting more hindcasts (we have tested up to 4 years either side) though skill may be underestimated
301 especially in these cases^{66,67}.

302 The overall procedure is as follows. For each start date i :

- 303 1. Compute the signal-adjusted (described above) NAO index of the ensemble mean $\hat{\text{NAO}}_i$
- 304 2. Compute the NAO index for each ensemble member NAO_i^k
- 305 3. For each ensemble member calculate the difference $\text{NAO}_i^k - \hat{\text{NAO}}_i$
- 306 4. Select the M ($= 20$) ensemble members with the smallest absolute differences

307 We take the mean of this subset of M members and present standardised forecast anomalies
308 (Figure 3) or adjust its variance to be the same as observed (Figure 2). We note that this approach
309 is applicable to forecasts as well as hindcasts. We present results for a subset of 20 members, but
310 the results are similar for subsets ranging from 10 to 40 members. This method relies on models

311 simulating realistic NAO teleconnections (F_{NAO}^k) and further improvements might be possible by
312 using the best models in this respect, but this is beyond the scope of this study.

313 **Significance.** For a given set of validation cases, we test for values that are unlikely to be ac-
314 counted for by uncertainties arising from a finite ensemble size (E) and a finite number of valida-
315 tion points (N). This is achieved using a non-parametric block bootstrap approach^{19,68,69}, in which
316 an additional 1000 hindcasts are created as follows:

- 317 1. Randomly sample with replacement N validation cases. In order to take autocorrelation into
318 account this is done in blocks of 5 consecutive cases.
- 319 2. For each of these, randomly sample with replacement E ensemble members.
- 320 3. Compute the required statistic for the ensemble mean (e.g. correlation, MSSS, RPC).
- 321 4. Repeat from (1) 1000 times to create a probability distribution.
- 322 5. Obtain the significance level based on a 2-tailed test of the hypothesis that skill is zero, or
323 RPC is one.

324 **Data Availability** The datasets analysed during the current study are available from the CMIP data archives.

325 **Code Availability** The code used during the current study is available from the corresponding author on
326 reasonable request.

327 **Acknowledgements** DMS, AAS, NJD, LH and RE were supported by the Met Office Hadley Centre
328 Climate Programme funded by BEIS and Defra and by the European Commission Horizon 2020 EUCP

329 project (GA 776613). FJDR, LPC, SW and RB also acknowledge the support from the EUCP project
330 (GA 776613) and from the Ministerio de Economía y Competitividad (MINECO) as part of the CLINSA
331 project (Grant No. CGL2017-85791-R). SW received funding from the innovation programme under the
332 Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433 and PO from the Ramon
333 y Cajal senior tenure programme of MINECO. The EC-Earth simulations were performed on Marenostrum
334 4 (hosted by the Barcelona Supercomputing Center, Spain) using Auto-Submit through computing hours
335 provided by PRACE. WAM, HP, KM and KP were supported by the German Federal Ministry for Education
336 and Research (BMBF) project MiKlip (grant 01LP1519A). NK, IB, FC and YW were supported by the
337 Norwegian Research Council projects SFE (grant 270733) the Nordic Center of excellent ARCPATH (grant
338 76654) and the Trond Mohn Foundation, under the project number : BFS2018TMT01 and received grants
339 for computer time from the Norwegian Program for supercomputing (NOTUR2, NN9039K) and storage
340 grants (NORSTORE, NS9039K). JM, LFB and DS are supported by Blue-Action (European Union Horizon
341 2020 research and innovation program, Grant Number: 727852) and EUCP (European Union Horizon 2020
342 research and innovation programme under grant agreement no 776613) projects. The National Center for
343 Atmospheric Research (NCAR) is a major facility sponsored by the US National Science Foundation (NSF)
344 under Cooperative Agreement No. 1852977. NCAR contribution was partially supported by the National
345 Oceanic and Atmospheric Administration (NOAA) Climate Program Office under Climate Variability and
346 Predictability Program Grant NA13OAR4310138 and by the US NSF Collaborative Research EaSM2 Grant
347 OCE-1243015.

348 **Competing Interests** The authors declare that there are no competing interests.

349 **Correspondence** Correspondence and requests for materials should be addressed to D.M.S.
350 (email: doug.smith@metoffice.gov.uk).

351 **Author contributions** D.M.S. led the analysis and writing with comments from all authors. R.E. pro-
352 cessed the CMIP5 data. A.A.S. suggested NAO-matching. All authors except A.A.S., P.A., A.B., P.-A.M.,
353 D.N., J.R. and P.R. contributed to creating the decadal prediction data.

- 354 1. Bindoff, N. L. *et al.* Detection and attribution of climate change: from global to regional. In
355 Stocker, T. F. *et al.* (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of*
356 *Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate*
357 *Change* (Cambridge University Press, 2013).
- 358
359 2. Kirtman, B. *et al.* Near-term climate change: Projections and predictability. In Stocker,
360 T. F. *et al.* (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working*
361 *Group I. to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*
362 (Cambridge University Press, 2013).
- 363 3. Collins, M. *et al.* Long-term climate change: Projections, commitments and irreversibility. In
364 Stocker, T. F. *et al.* (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of*
365 *Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate*
366 *Change*, 1029–1136 (Cambridge University Press, 2013).
- 367 4. Shepherd, T. G. Atmospheric circulation as a source of uncertainty in climate change projec-
368 tions. *Nature Geosci.* **7**, 703–708 (2014).
- 369 5. Hawkins, E. & Sutton, R. The potential to narrow uncertainty in projections of regional pre-
370 cipitation change. *Clim. Dyn.* **37**, 407–418 (2011).

- 371 6. Knutti, R. & Sedlek, J. Robustness and uncertainties in the new CMIP5 climate model projec-
372 tions. *Nature Climate Change* **3**, 369–373 (2013).
- 373 7. Hawkins, E., Smith, R. S., Gregory, J. M. & Stainforth, D. A. Irreducible uncertainty in
374 near-term climate projections. *Clim. Dyn.* **46**, 3807–3819 (2016).
- 375 8. Deser, C., Hurrell, J. W. & Phillips, A. S. The role of the North Atlantic Oscillation in Euro-
376 pean climate projections. *Clim. Dyn.* **49**, 3141–3157 (2017).
- 377 9. Marotzke, J. Quantifying the irreducible uncertainty in near term climate projections. *Wiley*
378 *Interdisciplinary Reviews: Climate Change* **10**, e563 (2019).
- 379 10. Scaife, A. A. & Smith, D. A signal-to-noise paradox in climate science. *npj Climate and*
380 *Atmospheric Science* **1**, 28 (2018).
- 381 11. Hawkins, E. & Sutton, R. The Potential to Narrow Uncertainty in Regional Climate Predic-
382 tions. *Bull. Am. Meteorol. Soc.* **90**, 1095–1108 (2009).
- 383 12. Fereday, D., Chadwick, R., Knight, J. & Scaife, A. A. Atmospheric Dynamics is the Largest
384 Source of Uncertainty in Future Winter European Rainfall. *J. Climate* **31**, 963–977 (2018).
- 385 13. Woollings, T. Dynamical influences on european climate: an uncertain future. *Philos. Trans.*
386 *R. Soc. London* **368**, 3733–3756 (2010).
- 387 14. Zappa, G. & Shepherd, T. G. Storylines of Atmospheric Circulation Change for European
388 Regional Climate Impact Assessment. *J. Climate* **30**, 6561–6577 (2017).

- 389 15. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment
390 design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
- 391 16. Boer, G. J. *et al.* The Decadal Climate Prediction Project (DCPP) contribution to CMIP6.
392 *Geosci. Model Devel.* (2016).
- 393 17. Hurrell, J. W., Kushnir, Y., Ottersen, G. & Visbeck, M. (eds.) *The North Atlantic Oscillation:
394 Climatic Significance and Environmental Impact*, vol. 134 of *Geophysical Monograph Series*
395 (American Geophysical Union, Washington, D. C., 2003).
- 396 18. Eade, R. *et al.* Do seasonal-to-decadal climate predictions underestimate the predictability of
397 the real world? *Geophys. Res. Lett.* **41**, 5620–5628 (2014).
- 398 19. Smith, D. M. *et al.* Robust skill of decadal climate predictions. *npj Climate and Atmospheric
399 Science* **2**, 13 (2019).
- 400 20. Siegert, S. *et al.* A Bayesian framework for verification and recalibration of ensemble fore-
401 casts: How uncertain is NAO predictability? *J. Climate* **29**, 995–1012 (2016).
- 402 21. Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and
403 precipitation. *Science* **269**, 676–679 (1995).
- 404 22. Scaife, A. A. *et al.* The CLIVAR C20C project: selected twentieth century climate events.
405 *Clim. Dyn.* **33**, 603–614 (2009).
- 406 23. Bracegirdle, T. J., Lu, H., Eade, R. & Woollings, T. Do CMIP5 Models Reproduce Observed
407 Low Frequency North Atlantic Jet Variability? *Geophys. Res. Lett.* **45**, 7204–7212 (2018).

- 408 24. Dobrynin, M. *et al.* Improved Teleconnection-Based Dynamical Seasonal Predictions of Bo-
409 real Winter. *Geophys. Res. Lett.* **45**, 3605–3614 (2018).
- 410 25. Simpson, I. R., Yeager, S. G., McKinnon, K. A. & Deser, C. Decadal predictability of late
411 winter precipitation in western Europe through an oceanjet stream connection. *Nature Geosci.*
412 **12**, 613–619 (2019).
- 413 26. Yeager, S. G. & Robson, J. I. Recent progress in understanding and predicting Atlantic decadal
414 climate variability. *Current Climate Change Reports* **3**, 112–127 (2017).
- 415 27. Eden, C. & Willebrand, J. Mechanism of interannual to decadal variability of the North At-
416 lantic circulation. *J. Climate* **14**, 2266–2280 (2001).
- 417 28. McCarthy, G. D., Haigh, I. D., Hirschi, J. J.-M., Grist, J. P. & Smeed, D. A. Ocean impact on
418 decadal Atlantic climate variability revealed by sea-level observations. *Nature* **521**, 508–510
419 (2015).
- 420 29. Clement, A. *et al.* The Atlantic Multidecadal Oscillation without a role for ocean circulation.
421 *Science* **350**, 320–324 (2015).
- 422 30. Zanardo, S., Nicotina, L., Hilberts, A. G. J. & Jewson, S. P. Modulation of Economic Losses
423 From European Floods by the North Atlantic Oscillation. *Geophys. Res. Lett.* **46**, 2563–2572
424 (2019).
- 425 31. Eden, C., Greatbatch, R. J. & Lu, J. Prospects for decadal prediction of the North Atlantic
426 Oscillation (NAO). *Geophys. Res. Lett.* **29**, 104–1–104–4 (2002).

- 427 32. Hoerling, M. P., Hurrell, J. W. & Xu, T. Tropical origins for recent North Atlantic climate
428 change. *Science* **292**, 90–92 (2001).
- 429 33. Greatbatch, R. J., Lin, H., Lu, J., Peterson, K. A. & Derome, J. Tropical/Extratropical forcing
430 of the AO/NAO: A corrigendum. *Geophys. Res. Lett.* **30** (2003).
- 431 34. Shin, S.-I. & Sardeshmukh, P. D. Critical influence of the pattern of Tropical Ocean warming
432 on remote climate trends. *Clim. Dyn.* **36**, 1577–1591 (2011).
- 433 35. Scaife, A. A. *et al.* Skillful long-range prediction of european and north american winters.
434 *Geophys. Res. Lett.* **41**, 2514–2519 (2014).
- 435 36. Dunstone, N. J. *et al.* Skilful seasonal predictions of summer European rainfall. *Geophys.*
436 *Res. Lett.* (2018).
- 437 37. Baker, L. H., Shaffrey, L. C., Sutton, R. T., Weisheimer, A. & Scaife, A. A. An intercomparison
438 of skill and over/underconfidence of the wintertime North Atlantic Oscillation in multi-model
439 seasonal forecasts. *Geophys. Res. Lett.* (2018).
- 440 38. Dunstone, N. J. *et al.* Skilful predictions of the winter North Atlantic Oscillation one year
441 ahead. *Nature Geosci.* (2016).
- 442 39. Yeager, S. G. *et al.* Predicting near-term changes in the earth system: A large ensemble of
443 initialized decadal prediction simulations using the Community Earth System Model. *Bull.*
444 *Am. Meteorol. Soc.* **99**, 1867–1886 (2018).

- 445 40. O'Reilly, C. H., Weisheimer, A., Woollings, T., Gray, L. J. & MacLeod, D. The importance of
446 stratospheric initial conditions for winter North Atlantic Oscillation predictability and impli-
447 cations for the signal-to-noise paradox. *Q. J. R. Meteorol. Soc.* **145**, 131–146 (2019).
- 448 41. Zhang, W. & Kirtman, B. Understanding the Signal-to-Noise Paradox with a Simple Markov
449 Model. *Geophys. Res. Lett.* 2019GL085159 (2019).
- 450 42. Jin, Y., Rong, X. & Liu, Z. Potential predictability and forecast skill in ensemble climate
451 forecast: a skill-persistence rule. *Clim. Dyn.* **51**, 2725–2741 (2018).
- 452 43. Strommen, K. & Palmer, T. N. Signal and noise in regime systems: A hypothesis on the
453 predictability of the North Atlantic Oscillation. *Q. J. R. Meteorol. Soc.* **145**, 147–163 (2019).
- 454 44. Czaja, A., Frankignoul, C., Minobe, S. & Vanni re, B. Simulating the Midlatitude Atmo-
455 spheric Circulation: What Might We Gain From High-Resolution Modeling of Air-Sea Inter-
456 actions? *Current Climate Change Reports* **5**, 390–406 (2019).
- 457 45. Scaife, A. A. *et al.* Does increased atmospheric resolution improve seasonal climate predic-
458 tions? *Atmos. Sci. Lett.* **20** (2019).
- 459 46. Mori, M., Kosaka, Y., Watanabe, M., Nakamura, H. & Kimoto, M. A reconciled estimate
460 of the influence of Arctic sea-ice loss on recent Eurasian cooling. *Nature Climate Change* **9**,
461 123–129 (2019).
- 462 47. Cheung, A. H. *et al.* Comparison of Low-Frequency Internal Climate Variability in CMIP5
463 Models and Observations. *J. Climate* **30**, 4763–4776 (2017).

- 464 48. Kravtsov, S. Pronounced differences between observed and CMIP5-simulated multidecadal
465 climate variability in the twentieth century. *Geophys. Res. Lett.* **44**, 5749–5757 (2017).
- 466 49. Wang, X., Li, J., Sun, C. & Liu, T. NAO and its relationship with the Northern Hemisphere
467 mean surface temperature in CMIP5 simulations. *J. Geophys. Res.* **122**, 4202–4227 (2017).
- 468 50. Kim, W. M., Yeager, S. G. & Danabasoglu, G. Key role of internal ocean dynamics in Atlantic
469 multidecadal variability during the last half century. *Geophys. Res. Lett.* **45** (2018).
- 470 51. Baker, A. J. *et al.* Enhanced Climate Change Response of Wintertime North Atlantic Circula-
471 tion, Cyclonic Activity, and Precipitation in a 25-km-Resolution Global Atmospheric Model.
472 *J. Climate* **32**, 7763–7781 (2019).
- 473 52. Morice, C. P., Kennedy, J. J., Rayner, N. A. & Jones, P. D. Quantifying uncertainties in
474 global and regional temperature change using an ensemble of observational estimates: The
475 HadCRUT4 data set. *J. Geophys. Res.* **117**, D08101 (2012).
- 476 53. Hansen, J., Ruedy, R., Sato, M. & Lo, K. Global surface temperature change. *Rev. Geophys.*
477 **48** (2010).
- 478 54. Karl, T. R. *et al.* Possible artifacts of data biases in the recent global surface warming hiatus.
479 *Science* **348**, 1469–1472 (2015).
- 480 55. Schneider, U. *et al.* GPCC’s new land surface precipitation climatology based on quality-
481 controlled in situ data and its role in quantifying the global water cycle. *Theor. Appl. Climatol.*
482 **115**, 15–40 (2014).

- 483 56. Allan, R. J. & Ansell, T. J. A new globally complete monthly historical gridded mean sea level
484 pressure data set (HadSLP2): 1850-2003. *J. Climate* **19**, 5816–5842 (2006).
- 485 57. Trenberth, K. E. & Shea, D. J. Atlantic hurricanes and natural variability in 2005. *Geophys.*
486 *Res. Lett.* **33**, L12704 (2006).
- 487 58. Doblas-Reyes, F. J. *et al.* Addressing model uncertainty in seasonal and annual dynamical
488 ensemble forecasts. *Q. J. R. Meteorol. Soc.* **135**, 1538–1559 (2009).
- 489 59. Hodson, D. L. R. & Sutton, R. T. Exploring multi-model atmospheric GCM ensembles with
490 ANOVA. *Climate Dynamics* **31**, 973–986 (2008).
- 491 60. Weisheimer, A. *et al.* How confident are predictability estimates of the winter North Atlantic
492 Oscillation? *Q. J. R. Meteorol. Soc.* **145**, 140–159 (2019).
- 493 61. Kumar, A. & Chen, M. Causes of skill in seasonal predictions of the Arctic Oscillation.
494 *Climate Dynamics* **51**, 2397–2411 (2018).
- 495 62. Borchert, L. F. *et al.* Decadal predictions of the probability of occurrence for warm summer
496 temperature extremes. *Geophys. Res. Lett.* (2019).
- 497 63. Krishnamurti, T. N. *et al.* Improved weather and seasonal climate forecasts from multimodel
498 superensemble. *Science* **285**, 1548–1550 (1999).
- 499 64. Yun, W. T., Stefanova, L. & Krishnamurti, T. N. Improvement of the multimodel superensem-
500 ble technique for seasonal forecasts. *J. Climate* **16**, 3834–3840 (2003).

- 501 65. Kug, J.-S., Lee, J.-Y., Kang, I.-S., Wang, B. & Park, C.-K. Optimal multi-model ensemble
502 method in seasonal prediction. *Asia-Pacific Journal of Atmospheric Sciences* **44**, 259–267
503 (2008).
- 504 66. Gangsto, R., Weigel, A. P., Lineger, M. A. & Appenzeller, C. Methodological aspects of the
505 validation of decadal predictions. *Climate Res.* **55**, 181–200 (2013).
- 506 67. Smith, D., Eade, R. & Pohlmann, H. A comparison of full-field and anomaly initialization for
507 seasonal to decadal climate prediction. *Clim. Dyn.* **41**, 3325–3338 (2013).
- 508 68. Wilks, D. S. *Statistical methods in the atmospheric sciences*, vol. 100 of *International geo-*
509 *physics series* (Academic Press, 2011), third edn.
- 510 69. Goddard, L. *et al.* A verification framework for interannual-to-decadal predictions experi-
511 ments. *Clim. Dyn.* **40**, 245–272 (2013).
- 512 70. Doblas-Reyes, F. J. *et al.* Using EC-Earth for climate prediction research. In *ECMWF Newslet-*
513 *ter* (ECMWF, 2018).
- 514 71. Haarsma, R. *et al.* HighResMIP versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR. De-
515 scription, model performance, data handling and validation. *Geosci. Model Dev.* (submitted).
- 516 72. Counillon, F. *et al.* Flow-dependent assimilation of sea surface temperature in isopycnal coor-
517 dinates with the Norwegian Climate Prediction Model. *Tellus A* **68**, 32437 (2016).
- 518 73. Wang, Y. *et al.* Optimising assimilation of hydrographic profiles into isopycnal ocean models
519 with ensemble data assimilation. *Ocean Modelling* **114**, 33–44 (2017).

- 520 74. Kharin, V. V., Boer, G. J., Merryfield, W. J., Scinocca, J. F. & Lee, W.-S. Statistical adjustment
521 of decadal predictions in a changing climate. *Geophys. Res. Lett.* **39**, L19705 (2012).
- 522 75. Swart, N. C. *et al.* The Canadian Earth System Model version 5 (CanESM5.0.3). *Geosci.*
523 *Model Devel.* **12**, 4823–4873 (2019).
- 524 76. Sospedra-Alfonso, R. & Boer, G. J. Assessing the impact of initialization on decadal prediction
525 skill. *Geophys. Res. Lett.* (2020).
- 526 77. Yang, X. *et al.* A predictable amo-like pattern in GFDL's fully-coupled ensemble initialization
527 and decadal forecasting system. *J. Climate* **26**, 650–661 (2013).
- 528 78. Williams, K. D. *et al.* The Met Office Global Coupled model 3.0 and 3.1 (GC3.0 and GC3.1)
529 configurations. *J. Adv. Model Earth Syst.* **10**, 357–380 (2018).
- 530 79. Müller, W. A. *et al.* Forecast skill of multi-year seasonal means in the decadal prediction
531 system of the Max Planck Institute for Meteorology. *Geophys. Res. Lett.* **39**, L22707 (2012).
- 532 80. Pohlmann, H. *et al.* Realistic Quasi-Biennial Oscillation Variability in Historical and Decadal
533 Hindcast Simulations Using CMIP6 Forcing. *Geophys. Res. Lett.* 2019GL084878 (2019).
- 534 81. Chikamoto, Y. *et al.* An overview of decadal climate predictability in a multi-model ensemble
535 by climate model MIROC. *Clim. Dyn.* **40**, 1201–1222 (2012).
- 536 82. Mochizuki, T. *et al.* Decadal prediction using a recent series of MIROC global climate models.
537 *J. Meteorol. Soc. Jpn* **90**, 373–383 (2012).

Table 1: Forecast systems and ensemble sizes.

Forecast Centre	Model	Atmosphere resolution ¹	Ocean resolution ²	Ensemble size	CMIP version
Barcelona Supercomputing Center, Spain	EC-Earth3 ^{70,71}	0.7x0.7x91x0.01	1x1x0.3x75	10	CMIP6
Bjerknes Center for Climate Research, Norway	NorCPM1 ^{72,73}	1.9x2.5x26x3	0.7x1.125x0.25x53	20	CMIP6
Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada	CanCM4 ⁷⁴	2.8x2.8x35x1	0.94x1.41x40	10	CMIP5
	CanESM5 ^{75,76}	2.8x2.8x49x1	1x1x0.3x45	10	CMIP6
Geophysical Fluid Dynamics Laboratory, USA	CM2.1 ⁷⁷	2x2.5x24x3	1x1x0.3x50	10	CMIP5
IPSL-EPOC, France	IPSL-CM6A-LR	1.25x2.5x79x0.005	1x1x0.3x75	10	CMIP6
Met Office Hadley Centre, UK	HadCM3 ⁶⁷	2.5x3.75x19x4.5	1.25x1.25x20	20	CMIP5
	HadGEM3 ⁷⁸	0.55x0.83x85x0.005	0.25x0.25x75	10	CMIP6
Max Planck Institute for Meteorology, Germany	MPI-ESM1.0-LR ⁷⁹	1.9x1.9x47x0.01	1.5x1.5x40	3	CMIP5
	MPI-ESM1.2-HR ⁸⁰	0.9x0.9x95x0.01	0.4x0.4x40	10	CMIP6
National Center for Atmospheric Research, USA	CESM1.1 ³⁹	0.9x1.25x30x2.26	1x1.125x0.27x60	40	CMIP6
University of Tokyo, National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology, Japan	MIROC5 ^{81,82}	1.4x1.4x40x3	1.4x1.4x0.5x49	6	CMIP5
	MIROC6	1.4x1.4x81x0.004	1x1x0.5x62	10	CMIP6

¹ Atmosphere resolution (degrees latitude)x(degrees longitude)x(number of vertical levels)x(lid height, hPa)

² Ocean resolution (degrees latitude)x(degrees longitude)x(optional degrees latitude at Equator)x(number of vertical levels)

538 **Figure 1** Decadal prediction skill for boreal winter (December to March) mean sea
539 level pressure. Skill for year 2-9 multi-model ensemble mean forecasts measured by (a)
540 anomaly correlation, (b) mean squared skill score (MSSS), (c) average anomaly correla-
541 tion for a 10-member ensemble mean (computed over 1000 random samples). (d) The
542 ratio of predictable components (RPC). RPC is not calculated where the correlation is
543 negative. Stippling shows where correlations and MSSS, or RPC, are significantly dif-
544 ferent to zero, or greater than one, respectively (95% confidence interval, see Methods).
545 Green boxes show the regions used to calculate the NAO.

546 **Figure 2** Underestimated signals. (a) Time series of observed (black curve) and model
547 forecast (years 2-9, red curve showing ensemble mean of 169 members and red shading
548 showing the 5-95% confidence interval diagnosed from the individual members) 8-year
549 running mean December to March NAO index. (b) As (a) but for ensemble mean forecast
550 rescaled to have the same variance as the observations (thin red curve), and addition-
551 ally smoothed by taking the lagged average of the latest four forecasts at each start date
552 (thick red curve, 676 members, see Methods). Forecast uncertainty (red shading, 5-95%
553 confidence interval) is obtained from the forecast ensemble mean error variance (see
554 Methods). (c) As (a) but for AMV and lagged ensemble. (d) As (c) but for NAO-matched
555 forecast (see Methods). (e, f) As (c, d) but for northern European rainfall. Values of
556 anomaly correlation (ACC) of the forecast ensemble mean and of persisting the latest

557 observed 8-year mean available before each start date, and the ratio of predictable com-
558 ponents (RPC), are indicated. Indices are defined in Methods. Time-series are anomalies
559 relative to the average of all year 2-9 hindcasts.

560 **Figure 3** Decadal predictions of the extreme NAO period (1986 to 1997). Observed
561 anomalies of (a) temperature, (b) precipitation and (c) mean sea level pressure. (d, e,
562 f) As (a, b, c) but for raw lagged ensemble mean forecasts. (g, h, i) As (d, e, f) but
563 standardised by the ensemble mean standard deviation. (j, k, l) As (d, e, f) but for NAO-
564 matched forecasts. Averages are taken for boreal winter (December to March) for all year
565 2-9 forecasts verifying in the period 1986 to 1997 (i.e. start dates 1985 to 1989 inclusive),
566 and converted to anomalies by removing the average over all hindcasts (i.e. start dates
567 1960 to 2005 inclusive). Units are standard deviations. The raw lagged ensemble (d, e,
568 f) is divided by the observed standard deviation to show the signal relative to observed
569 variability.

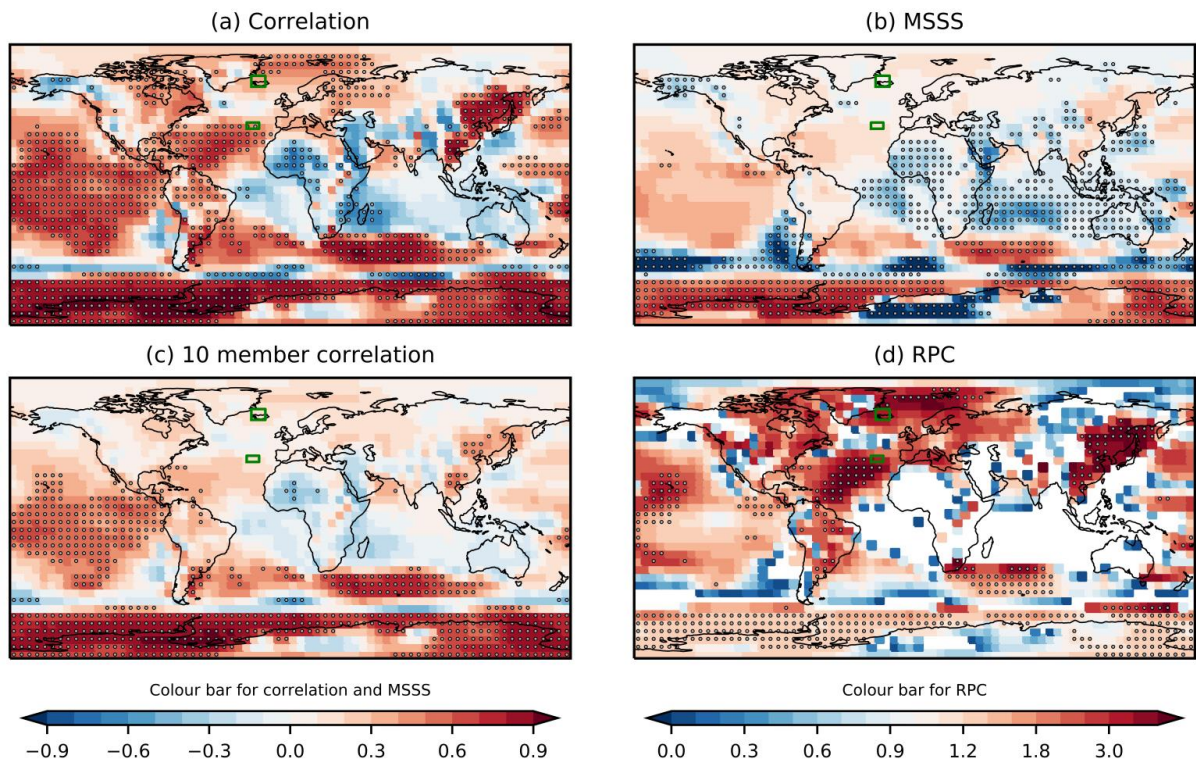


Figure 1: Decadal prediction skill for boreal winter (December to March) mean sea level pressure. Skill for year 2-9 multi-model ensemble mean forecasts measured by (a) anomaly correlation, (b) mean squared skill score (MSSS), (c) average anomaly correlation for a 10-member ensemble mean (computed over 1000 random samples). (d) The ratio of predictable components (RPC). RPC is not calculated where the correlation is negative. Stippling shows where correlations and MSSS, or RPC, are significantly different to zero, or greater than one, respectively (95% confidence interval, see Methods). Green boxes show the regions used to calculate the NAO.

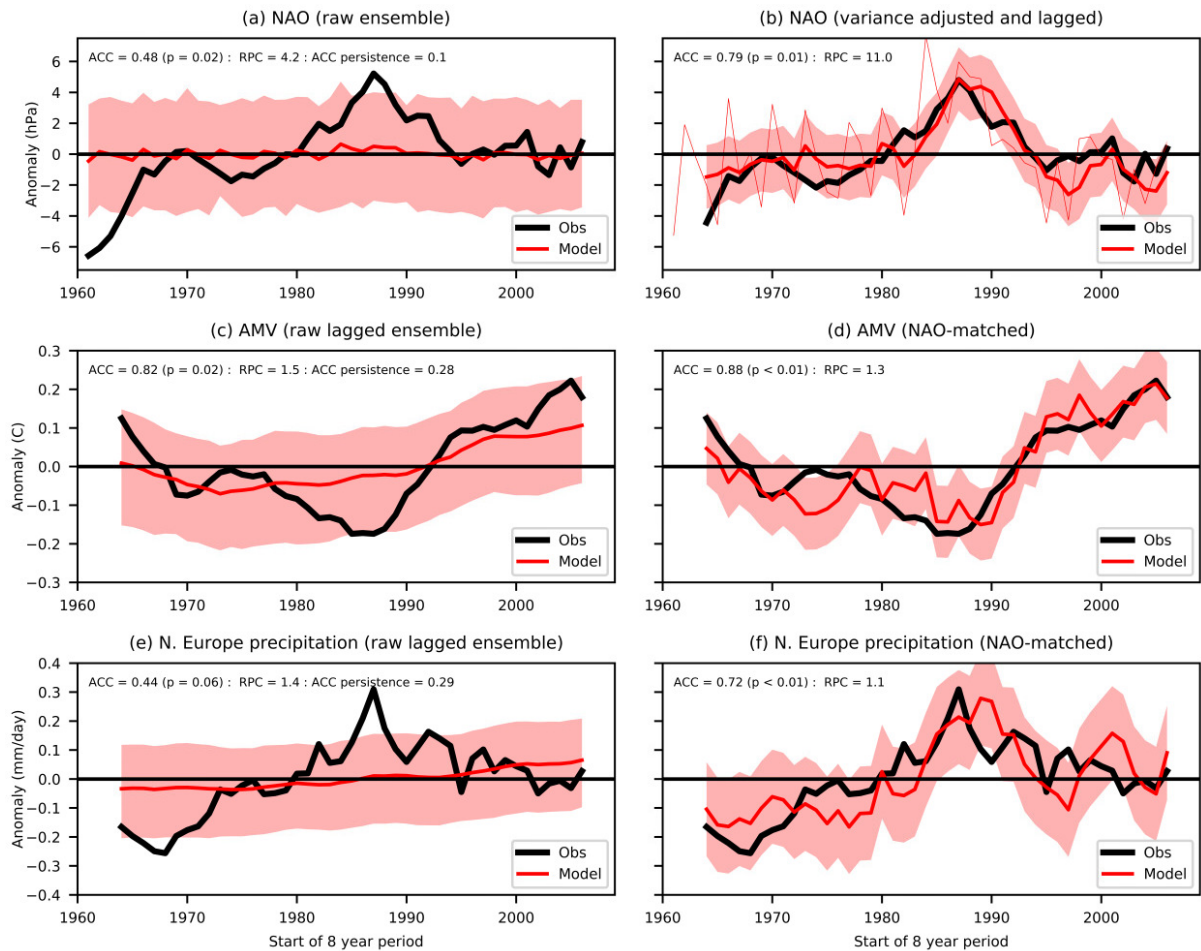


Figure 2: Underestimated signals. (a) Time series of observed (black curve) and model forecast (years 2-9, red curve showing ensemble mean of 169 members and red shading showing the 5-95% confidence interval diagnosed from the individual members) 8-year running mean December to March NAO index. (b) As (a) but for ensemble mean forecast rescaled to have the same variance as the observations (thin red curve), and additionally smoothed by taking the lagged average of the latest four forecasts at each start date (thick red curve, 676 members, see Methods). Forecast uncertainty (red shading, 5-95% confidence interval) is obtained from the forecast ensemble mean error variance (see Methods). (c) As (a) but for AMV and lagged ensemble. (d) As (c) but for NAO-matched forecast (see Methods). (e, f) As (c, d) but for northern European rainfall. Values of anomaly correlation (ACC) of the forecast ensemble mean and of persisting the latest observed 8-year mean available before each start date, and the ratio of predictable components (RPC), are indicated. Indices are defined in Methods. Time-series are anomalies relative to the average of all year 2-9 hindcasts.

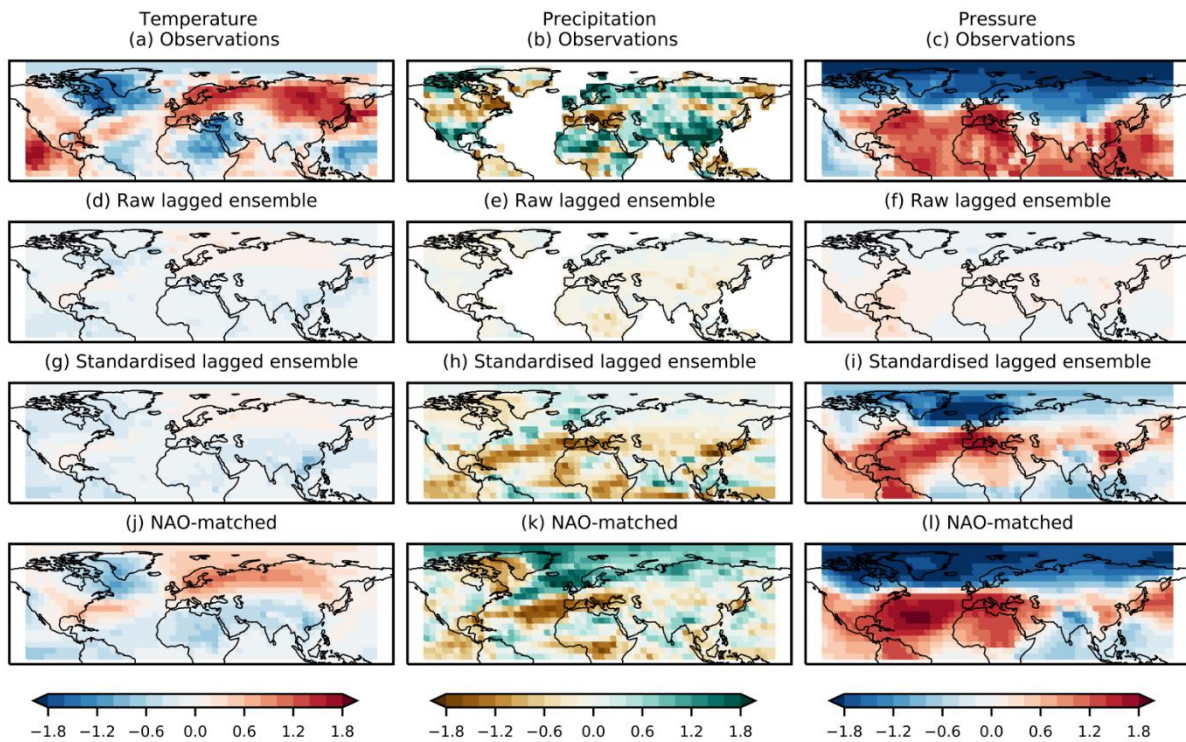


Figure 3: Decadal predictions of the extreme NAO period (1986 to 1997). Observed anomalies of (a) temperature, (b) precipitation and (c) mean sea level pressure. (d, e, f) As (a, b, c) but for raw lagged ensemble mean forecasts. (g, h, i) As (d, e, f) but standardised by the ensemble mean standard deviation. (j, k, l) As (d, e, f) but for NAO-matched forecasts. Averages are taken for boreal winter (December to March) for all year 2-9 forecasts verifying in the period 1986 to 1997 (i.e. start dates 1985 to 1989 inclusive), and converted to anomalies by removing the average over all hindcasts (i.e. start dates 1960 to 2005 inclusive). Units are standard deviations. The raw lagged ensemble (d, e, f) is divided by the observed standard deviation to show the signal relative to observed variability.