North Atlantic climate far more predictable than models imply

D. M. Smith[1], A. A. Scaife[1,2], R. Eade[1], P. Athanasiadis[3], A. Bellucci[3], I. Bethke[4], R. Bilbao[5], L. F. Borchert[6], L.-P. Caron[5], F. Counillon[4,7], G. Danabasoglu[8], T. Delworth[9], F. J. Doblas-Reyes[5,10], N. J. Dunstone[1], V. Estella-Perez[6], S. Flavoni[6], L. Hermanson[1], N. Keenlyside[4,7], V. Kharin[11], M. Kimoto[12], W. J. Merryfield[11], J. Mignot[6], T. Mochizuki[13,14], K. Modali[15], P.-A. Monerie[16], W. A. Müller[15], D. Nicolí[3], P. Ortega[5], K. Pankatz[17], H. Pohlmann[15,17], J. Robson[16], P. Ruggieri[3], R. Sospedra-Alfonso[11], D. Swingedouw[18],Y. Wang[7], S. Wild[5], S. Yeager[8], X. Yang[9] and L. Zhang[9]

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

[5]*Barcelona Supercomputing Center, Jordi Girona 29 - 08034 Barcelona, Spain*

[6]*Sorbonne Universités, LOCEAN Laboratory, Institut Pierre Simon Laplace (IPSL), Paris, France*

[7]*Nansen Environmental and Remote Sensing Center, and Bjerknes Centre for Climate Research, Bergen, Norway*

[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,*

Victoria, British Columbia, Canada

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

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

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

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

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

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

[18]*CNRS-EPOC, Université de Bordeaux, Pessac, France*

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

Abstract

Quantifying signals and uncertainties in climate models is essential for climate change detection, attribution, prediction and projection[1–3]. Although inter-model agreement is high for large-scale temperature signals, dynamical changes in atmospheric circulation are very uncertain[4], leading to low confidence in regional projections especially for precipitation over the coming decades[5,6]. Furthermore, model simulations with tiny differences in initial conditions suggest that uncertainties may be largely irreducible due to the chaotic nature of the climate system[7–9]. However, climate projections are difficult to verify until further observations become available. Here we assess retrospective climate model predictions of the

last six decades and show that decadal variations in north Atlantic winter climate are highly predictable despite a lack of agreement between individual model simulations and little predictive ability of raw model outputs. Crucially, models underestimate the predictable signal of the North Atlantic Oscillation (NAO, the leading mode of north Atlantic atmospheric circulation variability) by an order of magnitude. Consequently, compared to perfect models, 100 times more ensemble members are needed to extract the NAO signal, and its climate impacts are underestimated relative to other factors. To address these limitations, we implement a two-stage post-processing technique that first takes the variance-adjusted ensemble mean NAO and then selects the ensemble members with the required NAO signal. This approach yields skilful decadal predictions of European and eastern North American winters. Atlantic Multidecadal variability is also improved, suggesting skill does not arise solely from the north Atlantic Ocean. Our results highlight the pressing need to understand why the signal-to-noise ratio is too small in climate models[10], and the extent to which correcting this model error would reduce uncertainties in regional climate change on timescales beyond a decade.

Global climate models are used extensively to understand the drivers of past climate variability and change, and to predict what is likely to happen in the future[1–3]. Underpinning this is a need for accurate estimates of signals and associated uncertainties in climate model simulations in order to distinguish between different causes of past climate change, and to provide reliable confidence limits on future projections. Uncertainties are typically partitioned into three sources[11]: scenario uncertainty arising from an imperfect knowledge of external forcing factors, including changes in greenhouse gases, ozone, anthropogenic and volcanic aerosols, and solar irradiance; modelling uncertainty arising from the fact that different models respond differently to the same radiative forcing; and internal variability of climate that would occur in the absence of any external forcing.

Climate projections for many regions are currently highly uncertain, especially for atmospheric circulation[4,12] and related impacts, including precipitation[5,6]. This is particularly well illustrated by the fact that modelling[13,14] and internal variability[7,8] uncertainties are each large enough to allow opposite projections of European winters, especially for the coming decades. Whilst modelling uncertainties might be reduced as models improve, internal variability uncertainties have been interpreted to be largely irreducible[7–9] suggesting that confident projections of European winters may never be possible. However, such conclusions assume that signals and uncertainties diagnosed from climate models are correct. Although multi-decadal and longer climate projections are difficult to verify until future observations become available, signals over the first 10 years can be more robustly evaluated using retrospective decadal predictions (hereafter referred to as hindcasts).

We use a very large multi-model ensemble of decadal hindcasts from the Coupled Model Intercomparison Project (CMIP) phases 5[15] and 6[16]. We focus on the boreal winter period (December to March) averaged over forecast years 2 to 9 to avoid seasonal to annual predictability and focus on decadal timescales. We use hindcasts starting each year over the period 1960 to 2005 from 6 CMIP5 and 8 CMIP6 modelling systems, giving a total of 169 ensemble members which are weighted equally (see Methods, Table 1). Hence our total hindcast dataset comprises 77,740

(46 start dates times 169 ensemble members times 10 years) years of model integrations to provide robust statistics.

To compare with uncertainties in climate projections[5,7,8,13,14] we focus on European winters which are largely controlled by the North Atlantic Oscillation (NAO), the leading mode of atmospheric circulation variability in the north Atlantic[17]. The NAO represents the meridional gradient in mean sea level pressure (mslp), typically measured as the difference in pressure between the Azores and Iceland. Its positive (negative) phase reflects an increased (reduced) pressure gradient driving stronger (weaker) mid-latitude westerly winds with increased (reduced) storminess, and a northward (southward) shift of the jet stream. Impacts of the NAO are characterised by a quadrupole pattern, with a positive (negative) NAO driving warmer, wetter (colder, drier) conditions in northern Europe and south-east North America along with colder, drier (warmer, wetter) conditions in southern Europe and north-east North America.

We assess skill using two different measures (see Methods): anomaly correlation coefficient (ACC) which measures the phase of variability, and mean-squared-skill-score (MSSS) which measures the amplitude of variability. We find significant skill for decadal predictions of winter mslp in most regions, including the north Atlantic, when measured by the ACC between the 169-member ensemble mean and observations (Figure 1a). However, skill is much lower especially in the north Atlantic when measured by the MSSS or the ACC of a smaller (10-member, typical of individual prediction systems[16]) ensemble mean (Figure 1 b and c). Timeseries from the observations and each model ensemble member consist of a predictable component (the signal) and unpredictable internal variability (the noise). The discrepancy in skill between ACC and MSSS, and the need for a large ensemble, arise because the signal-to-noise ratio is too small in the models compared to observations[10,18,19]. Hence, skill is low in a 10-member ensemble mean because a larger ensemble is required to reduce the noise and extract the predicted signal. In contrast, the signal resulting from a large ensemble mean may capture the correct phase of observed variability giving a significant ACC, but its amplitude will be much too small resulting in a low MSSS.

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

The NAO exhibits marked decadal variability[21] with a strong increase from the 1960s to the 1990s and a decrease thereafter (Figure 2a, black curve). The raw ensemble mean forecast shows virtually no signal (Figure 2a, red curve), and the observations generally lie within the model uncertainties (shading showing the 5-95% range diagnosed from the ensemble spread), although the extreme values in the early 1960s and late 1980s are not well-captured by models in agreement with other studies[22,23]. Taken at face value, as is done for climate projections[5,7,8,14], the small model signal and much larger spread would imply little ability to predict the NAO and a large component of unpredictable internal variability. However, by comparing with observations we find significant correlation skill of the ensemble mean (ACC=0.48, p=0.02), while persistence provides a poor forecast (ACC=0.1). Hence, skilful climate model predictions of the NAO are possible using the ensemble mean, but the signal-to-noise ratio is too small (RPC=4.2) and its variance must be calibrated to provide realistic forecasts[19].

Rescaling the ensemble mean time-series to have the same variance as the observations reveals that the predictions do capture the observed increase from the 1960s to 1990s and decrease thereafter (Figure 2b). However, even with 169 ensemble members (Figure 2b thin red curve) there are large interannual variations that are not expected or observed in 8-year rolling means. We therefore create a larger lagged ensemble by taking the average of the four latest forecasts available at each start date (giving 676 members, Figure 2b thick red curve, see Methods). This reveals that the NAO is highly predictable on decadal timescales (ACC=0.79, $p<0.01$) in stark contrast to the lack of predictability implied by the standard interpretation of raw model output (Figure 2a). Importantly, the signal-to-noise ratio is much too small in the models (RPC=11, p=0.02). The total 8-year variability of the NAO in individual model members (standard deviation = 1.7 to 2.6 hPa, 5-95% range, year 2-9 hindcasts) is not significantly different to the observations (2.4hPa). Hence the predictable signal (see Methods) is underestimated by an order of magnitude in the model ensemble. Since the standard error of the ensemble mean is reduced by the square root of the ensemble size, the ensemble required to extract the signal is 100 times larger than it would be for perfect models.

The fact that the NAO signal is much too weak in models implies that the impacts of the

NAO will be underestimated relative to other factors such as greenhouse gases. Hence in regions influenced by the NAO the ensemble mean will not reflect the true balance of driving factors and simply inflating its variance to be the same as observed will not correct the error. A potential solution is to post-process the model output by selecting a subset of (20) ensemble members from the lagged ensemble (of 676 members) whose simulated NAO is closest in sign and magnitude to the ensemble mean NAO after adjusting this to take into account the underestimated signal. These members contain close-to the correct magnitude of the forecast NAO whilst retaining influences from greenhouse gases and other sources. We refer to this procedure as "NAO-matching" (see Methods) and note that it builds on previous techniques[24,25] by using the models as much as possible instead of observed relationships which may not be causal or robust.

We investigate this technique first for forecasts of Atlantic Multidecadal Variability (AMV, see Methods). AMV is thought to be one of the most predictable aspects of decadal climate[26], yet the lagged ensemble mean does not capture the correct timing of the minimum in the late 1980s (Figure 2c). NAO-matching captures the minimum and subsequent rapid warming in much better agreement with observations (Figure 2d) consistent with evidence that AMV is at least partly forced by the NAO[27–29]. We find similar improvements for northern European rainfall: the lagged ensemble mean is not significantly skilful and the observations lie outside the modelled uncertainties in the 1960s and 1980s (Figure 2e), whereas the NAO-matched ensemble is significantly skilful (ACC=0.72, $p<0.01$) and captures the observed increase from the 1960s to late 1980s and decrease thereafter. As expected, these improvements are not seen by simply adjusting the variance of the ensemble mean (Extended Data Figure 1).

Forecasts of extreme decades would be of particular value since they could enable action to be taken in advance to avoid the most severe climate impacts[30]. We therefore investigate the extreme positive NAO period between 1986 and 1997 (8-year means starting 1986 to 1990, Figure 2a). Consistent with the above results, the raw lagged ensemble mean shows virtually no signal compared to observed variability (Figure 3 a, b, c compared to d, e, f). Adjusting its variance to be equal to the observed variance (Figure 3 g, h, i) reveals that the forecasts do capture the positive NAO (as expected from Figure 2b), but the expected impacts are underestimated, especially for temperature and northern European precipitation. However, the NAO-matched forecast (Figure 3 j, k, l) shows a clear improvement and captures the expected quadrupole pattern with warm, wet (cold, dry) anomalies in northern Eurasia and south-east North America (northern Africa and parts of southern Europe, and north-east North America), as well as low pressure across the Arctic. Similar improvements from NAO-matching are found for trends and for skill measured over all of the hindcasts (Extended Data Figures 2 to 4).

We have shown that the winter NAO and related impacts on Europe and eastern North America are highly predictable on decadal timescales. AMV is usually believed to be a major source of decadal prediction skill[26,31]. However, we find that predictions of AMV can be improved by using the forecast NAO (Figure 2c,d), whereas predictions of the NAO are degraded by selecting the most skilful AMV ensemble members (Extended Data Figure 5). This suggests that the NAO is not solely driven by AMV. Hence other potential influences, including for example the tropics[32–34], warrant further investigation.

Crucially we find that the NAO signal is underestimated by an order of magnitude in the model ensemble. This adds to an increasing body of evidence that the signal-to-noise ratio is too small in climate models, seen on seasonal[20,35–37], interannual[38] and decadal[19,39,40] timescales. Consequently, the real world is more predictable than climate models suggest[10,18] and uncertainties diagnosed from raw model simulations are too large. The cause of this error is not yet known, though there are several hypotheses including weak teleconnections to the quasi-biennial oscillation[41], lack of persistence in the NAO[42,43] and in weather regimes[44], unresolved ocean atmosphere interactions[45] and weak transient eddy feedback[46].

A key question is whether climate models also underestimate signals on timescales beyond a decade. There is some evidence that the atmospheric circulation response to Arctic sea ice loss[47], and to external factors[10] including volcanic eruptions, solar variations and ozone changes, are too weak in models. Models also appear to underestimate the magnitude of multi-decadal temperature variability[48,49] especially for the north Atlantic[50,51]. Furthermore, model-simulated winter climate change signals in the north Atlantic increase substantially as resolution increases[52], consistent with the suggestion that eddy feedbacks are inadequately resolved[46]. If this is robust, treating current model simulations at face value is giving misleading conclusions about uncertainties and irreducible internal variability.

References

1. Bindoff, N. L. *et al.* Detection and attribution of climate change: from global to regional. In Stocker, T. F. *et al.* (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2013).

2. Kirtman, B. *et al.* Near-term climate change: Projections and predictability. In Stocker, T. F. *et al.* (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I. to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2013).

3. Collins, M. *et al.* Long-term climate change: Projections, commitments and irreversibility. In Stocker, T. F. *et al.* (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, 1029–1136 (Cambridge University Press, 2013).

4. Shepherd, T. G. Atmospheric circulation as a source of uncertainty in climate change projections. *Nature Geosci.* **7**, 703–708 (2014).

5. Hawkins, E. & Sutton, R. The potential to narrow uncertainty in projections of regional precipitation change. *Clim. Dyn.* **37**, 407–418 (2011).

6. Knutti, R. & Sedlek, J. Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change* **3**, 369–373 (2013).

7. Hawkins, E., Smith, R. S., Gregory, J. M. & Stainforth, D. A. Irreducible uncertainty in near-term climate projections. *Clim. Dyn.* **46**, 3807–3819 (2016).

8. Deser, C., Hurrell, J. W. & Phillips, A. S. The role of the North Atlantic Oscillation in European climate projections. *Clim. Dyn.* **49**, 3141–3157 (2017).

9. Marotzke, J. Quantifying the irreducible uncertainty in near term climate projections. *Wiley Interdisciplinary Reviews: Climate Change* **10**, e563 (2019).

10. Scaife, A. A. & Smith, D. A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science* **1**, 28 (2018).

11. Hawkins, E. & Sutton, R. The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bull. Am. Meteorol. Soc.* **90**, 1095–1108 (2009).

12. Fereday, D., Chadwick, R., Knight, J. & Scaife, A. A. Atmospheric Dynamics is the Largest Source of Uncertainty in Future Winter European Rainfall. *J. Climate* **31**, 963–977 (2018).

13. Woollings, T. Dynamical influences on european climate: an uncertain future. *Philos. Trans. R. Soc. London* **368**, 3733–3756 (2010).

14. Zappa, G. & Shepherd, T. G. Storylines of Atmospheric Circulation Change for European Regional Climate Impact Assessment. *J. Climate* **30**, 6561–6577 (2017).

15. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).

16. Boer, G. J. *et al.* The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geosci. Model Devel.* (2016).

17. Hurrell, J. W., Kushnir, Y., Ottersen, G. & Visbeck, M. (eds.) *The North Atlantic Oscillation: Climatic Significance and Environmental Impact*, vol. 134 of *Geophysical Monograph Series* (American Geophysical Union, Washington, D. C., 2003).

18. Eade, R. *et al.* Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.* **41**, 5620–5628 (2014).

19. Smith, D. M. *et al.* Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science* **2**, 13 (2019).

20. Siegert, S. *et al.* A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *J. Climate* **29**, 995–1012 (2016).

21. Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).

22. Scaife, A. A. *et al.* The CLIVAR C20C project: selected twentieth century climate events. *Clim. Dyn.* **33**, 603–614 (2009).

23. Bracegirdle, T. J., Lu, H., Eade, R. & Woollings, T. Do CMIP5 Models Reproduce Observed Low Frequency North Atlantic Jet Variability? *Geophys. Res. Lett.* **45**, 7204–7212 (2018).

24. Dobrynin, M. *et al.* Improved Teleconnection-Based Dynamical Seasonal Predictions of Boreal Winter. *Geophys. Res. Lett.* **45**, 3605–3614 (2018).

25. Simpson, I. R., Yeager, S. G., McKinnon, K. A. & Deser, C. Decadal predictability of late winter precipitation in western Europe through an ocean-jet stream connection. *Nature Geosci.* **12**, 613–619 (2019).

26. Yeager, S. G. & Robson, J. I. Recent progress in understanding and predicting Atlantic decadal climate variability. *Current Climate Change Reports* **3**, 112–127 (2017).

27. Eden, C. & Willebrand, J. Mechanism of interannual to decadal variability of the North Atlantic circulation. *J. Climate* **14**, 2266–2280 (2001).

28. McCarthy, G. D., Haigh, I. D., Hirschi, J. J.-M., Grist, J. P. & Smeed, D. A. Ocean impact on decadal Atlantic climate variability revealed by sea-level observations. *Nature* **521**, 508–510 (2015).

29. Clement, A. *et al.* The Atlantic Multidecadal Oscillation without a role for ocean circulation. *Science* **350**, 320–324 (2015).

30. Zanardo, S., Nicotina, L., Hilberts, A. G. J. & Jewson, S. P. Modulation of Economic Losses From European Floods by the North Atlantic Oscillation. *Geophys. Res. Lett.* **46**, 2563–2572 (2019).

31. Eden, C., Greatbatch, R. J. & Lu, J. Prospects for decadal prediction of the North Atlantic Oscillation (NAO). *Geophys. Res. Lett.* **29**, 104–1–104–4 (2002).

32. Hoerling, M. P., Hurrell, J. W. & Xu, T. Tropical origins for recent North Atlantic climate change. *Science* **292**, 90–92 (2001).

33. Greatbatch, R. J., Lin, H., Lu, J., Peterson, K. A. & Derome, J. Tropical/Extratropical forcing of the AO/NAO: A corrigendum. *Geophys. Res. Lett.* **30** (2003).

34. Shin, S.-I. & Sardeshmukh, P. D. Critical influence of the pattern of Tropical Ocean warming on remote climate trends. *Clim. Dyn.* **36**, 1577–1591 (2011).

35. Scaife, A. A. *et al.* Skillful long-range prediction of european and north american winters. *Geophys. Res. Lett.* **41**, 2514–2519 (2014).

36. Dunstone, N. J. *et al.* Skilful seasonal predictions of summer European rainfall. *Geophys. Res. Lett.* (2018).

37. Baker, L. H., Shaffrey, L. C., Sutton, R. T., Weisheimer, A. & Scaife, A. A. An intercomparison of skill and over/underconfidence of the wintertime North Atlantic Oscillation in multi-model seasonal forecasts. *Geophys. Res. Lett.* (2018).

38. Dunstone, N. J. *et al.* Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nature Geosci.* (2016).

39. Yeager, S. G. *et al.* Predicting near-term changes in the earth system: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bull. Am. Meteorol. Soc.* **99**, 1867–1886 (2018).

40. Athanasiadis, P. J. *et al.* Decadal predictability of North Atlantic blocking and the NAO. *npj Climate and Atmospheric Science* **3**, 20 (2020).

41. O'Reilly, C. H., Weisheimer, A., Woollings, T., Gray, L. J. & MacLeod, D. The importance of stratospheric initial conditions for winter North Atlantic Oscillation predictability and implications for the signal-to-noise paradox. *Q. J. R. Meteorol. Soc.* **145**, 131–146 (2019).

42. Zhang, W. & Kirtman, B. Understanding the Signal-to-Noise Paradox with a Simple Markov Model. *Geophys. Res. Lett.* 2019GL085159 (2019).

43. Jin, Y., Rong, X. & Liu, Z. Potential predictability and forecast skill in ensemble climate forecast: a skill-persistence rule. *Clim. Dyn.* **51**, 2725–2741 (2018).

44. Strommen, K. & Palmer, T. N. Signal and noise in regime systems: A hypothesis on the predictability of the North Atlantic Oscillation. *Q. J. R. Meteorol. Soc.* **145**, 147–163 (2019).

45. Czaja, A., Frankignoul, C., Minobe, S. & Vannière, B. Simulating the Midlatitude Atmospheric Circulation: What Might We Gain From High-Resolution Modeling of Air-Sea Interactions? *Current Climate Change Reports* **5**, 390–406 (2019).

46. Scaife, A. A. *et al.* Does increased atmospheric resolution improve seasonal climate predictions? *Atmos. Sci. Lett.* **20** (2019).

47. Mori, M., Kosaka, Y., Watanabe, M., Nakamura, H. & Kimoto, M. A reconciled estimate of the influence of Arctic sea-ice loss on recent Eurasian cooling. *Nature Climate Change* **9**, 123–129 (2019).

48. Cheung, A. H. *et al.* Comparison of Low-Frequency Internal Climate Variability in CMIP5 Models and Observations. *J. Climate* **30**, 4763–4776 (2017).

49. Kravtsov, S. Pronounced differences between observed and CMIP5-simulated multidecadal climate variability in the twentieth century. *Geophys. Res. Lett.* **44**, 5749–5757 (2017).

50. Wang, X., Li, J., Sun, C. & Liu, T. NAO and its relationship with the Northern Hemisphere mean surface temperature in CMIP5 simulations. *J. Geophys. Res.* **122**, 4202–4227 (2017).

51. Kim, W. M., Yeager, S. G. & Danabasoglu, G. Key role of internal ocean dynamics in Atlantic multidecadal variability during the last half century. *Geophys. Res. Lett.* **45** (2018).

52. Baker, A. J. *et al.* Enhanced Climate Change Response of Wintertime North Atlantic Circulation, Cyclonic Activity, and Precipitation in a 25-km-Resolution Global Atmospheric Model. *J. Climate* **32**, 7763–7781 (2019).

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)

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.

Methods

Observations and models. Near surface temperature observations are computed as the average of HadCRUT4[52], NASA-GISS[53] and NCDC[54]. Precipitation observations are taken from GPCC[55] and mean sea level pressure is taken from HadSLP2[56].

We assess a large multi-model ensemble (169 members, Table 1) of decadal predictions from 14 modelling systems using hindcasts starting each year from 1960 to 2005 from the Coupled Model Intercomparison Project (CMIP5) phases 5[15] and 6[16]. We found no significant difference in NAO correlation skill between the CMIP5 and CMIP6 ensembles and focus on the combined ensemble to obtain the most robust statistics. We create ensemble means by taking the equally-weighted average of all ensemble members and assess rolling 8-year boreal winter (December to March) means defined by calendar years 2 to 9 from each start date. The forecasting systems start between 1st of November and January each year, giving a lead time of at least a year before the assessed forecast period to focus on decadal timescales and avoid predictability arising from seasonal to annual variability. Both halves of the 8-year period contribute to skill (NAO ACC = 0.57 and 0.45, p=0.03, for forecast years 2 to 5 and 6 to 9 respectively). Both observations and models were interpolated to a 5° longitude by 5° latitude grid before comparison.

Indices. The North Atlantic Oscillation (NAO) index is calculated as the difference in mean sea level pressure between two small boxes located around the Azores (28-20°W, 36-40°N) and Iceland (25-16°W, 63-70°N) with the average over the whole time series removed to create anomalies[38]. Atlantic Multidecadal Variability (AMV) is calculated as the near-surface temperature in the North

Atlantic (80-0°W, 0-60°N) minus the global average (60°S-60°N)[57]. European rainfall is averaged over the box 10°W-25°E, 55-70°N. All forecasts indices are based on the ensemble mean.

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

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

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

$$\text{Ratio of predictable components RPC } = \frac{\sigma^o_{sig}/\sigma^o_{tot}}{\sigma^f_{sig}/\sigma^f_{tot}} = \frac{ACC}{\sigma^f_{sig}/\sigma^f_{tot}} \tag{3}$$

$$\text{Ratio of predictable signals } = \frac{\sigma^o_{sig}}{\sigma^f_{sig}} = RPC\frac{\sigma^o_{tot}}{\sigma^f_{tot}} \tag{4}$$

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

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

Uncertainties in raw model forecasts are computed from the ensemble standard deviation for each start date. Uncertainties in variance adjusted and NAO-matched forecasts are computed from the root-mean-square error between the ensemble mean and the observations as required for reliable forecasts[58].

We note that it is theoretically possible for the multi-model RPC to be larger than for individual models if time dependent model biases[59] or teleconnection errors reduce the model signal more than the correlation with observations. Assessing this thoroughly would require large ensembles of individual model hindcasts which are not available. However, assessing the largest individual model ensemble available (NCAR CESM1.1 with 40 members per year, giving 160 lagged members, Table 1) does not support this hypothesis: the NCAR RPC of 6.2 is not significantly different from the average RPC of multi-model ensembles of the same size (4.8 averaged over 1000 random samples, with 5-95% range 1.3 to 7.4). Furthermore, the statistics presented in this study are appropriate for multi-model ensemble forecasts.

We further note that there is some evidence that the predictability of the NAO may vary on multi-decadal timescales[60], though this is not robust across models[61]. Our results are statistically significant for the hindcast period available, but longer hindcasts that include more cycles of decadal variability would be beneficial for future studies.

Lagged ensemble. Consecutive 8-year means contain 7 identical years. Hence large interannual variations, as seen in 169-member ensemble mean NAO forecsts (Figure 2b), are not expected.

They occur because the signal to noise ratio is too small in models and consecutive decadal predictions consist of independent model simulations that are dominated by different samples of the noise. Ideally additional ensemble members would be used to reduce the noise further, but these are not available. Instead we create a lagged ensemble by combining the required forecast with the previous three i.e. the year 2-9 forecasts starting in 1963 are combined with the year 2-9 forecasts starting in 1962, 1961 and 1960 giving a total of 676 members (169 members time 4 start dates). The previous forecasts are sub-optimal because they do not cover exactly the same forecast period, and rely on the persistence of running 8-year means. Hence there is a trade off between reducing the noise with additional members and potentially degrading the skill by relying on persistence. In the current generation of climate models the benefit in reducing the noise far outweighs the degradation from using persistence. We present results for the combination of 4 lagged forecasts, but find similar levels of skill for other combinations (NAO ACC = 0.71 and 0.78 for combining 3 and 5 lagged forecasts respectively). A similar technique relying on persistence of the predictor recently proved to strongly reduce the noise in decadal predictions of summer temperature extremes over land[62].

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

$$O = O_{NAO} + O_{OTHER} + \epsilon^o \tag{5}$$

$$F^k = F^k_{NAO} + F^k_{OTHER} + \epsilon^k \tag{6}$$

$$\hat{F} = \hat{F}_{NAO} + \hat{F}_{OTHER} + \hat{\epsilon} \tag{7}$$

where O, F^k and $\hat{F}$ are the observed, forecast ensemble member *k* and forecast ensemble mean values of a meteorological variable (e.g. temperature, rainfall, pressure). The subscript *NAO* refers to the portion that is related to the NAO, the subscript *OTHER* refers to the portion related to other predictable drivers (including greenhouse gases and sea surface temperatures unrelated to the NAO) and ϵ is an unpredictable residual. Because the predictable NAO signal is too small in models, the mean of a very large ensemble is required for skilful NAO predictions (Figure 2b). However, the magnitude of the ensemble mean NAO is much too small (Figure 2a) and therefore $\hat{F}_{NAO}$ will be severely underestimated.

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

An alternative approach[24] replaces the underestimated $\hat{F}_{NAO}$ with more realistic F^k_{NAO} by selecting from the full ensemble a smaller set of members that have the required magnitude of the NAO. These members contain close-to the correct magnitude of the required NAO and its teleconnections whilst retaining other influencies. Hence, $\hat{F}_{NAO}$ for this selected ensemble will be larger than that of the full ensemble, thereby increasing the signal. Because the selected ensemble is smaller the remaining noise will not be reduced as much as in the full ensemble. However, the selection process transfers variability from what would be considered as noise in a random ensemble into $\hat{F}_{NAO}$, thereby reducing $\hat{\epsilon}$ in the selected ensemble. Hence, in regions affected by the NAO the increase in signal is likely to be larger than the reduced suppression of the remaining noise, thereby increasing the signal to noise ratio and improving the skill.

In the previous seasonal forecast study[24] the required NAO was obtained based on observed relationships with potential drivers. However, on decadal timescales such relationships are not well-established and are more likely to be affected by sampling errors. We therefore take the required NAO to be the ensemble mean forecast NAO but adjusted to account for the underestimation of the predictable signal. This is achieved by muliplying the ensemble mean NAO by the ratio of predictable signals (equation 4). To avoid overfitting to observations we compute the ratio of predictable signals for each hindcast start date separately using a cross-validation approach in which the required hindcast and those on either side are omitted. Our conclusions are robust to omitting more hindcasts (we have tested up to 4 years either side) though skill may be underestimated especially in these cases[66,67].

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

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

We take the mean of this subset of M members and present standardised forecast anomalies (Figure 3) or adjust its variance to be the same as observed (Figure 2). We note that this approach is applicable to forecasts as well as hindcasts. We present results for a subset of 20 members, but the results are similar for subsets ranging from 10 to 40 members. This method relies on models simulating realistic NAO teleconnections (F^k_{NAO}) and further improvements might be possible by using the best models in this respect, but this is beyond the scope of this study.

Significance. For a given set of validation cases, we test for values that are unlikely to be accounted for by uncertainties arising from a finite ensemble size (E) and a finite number of validation points (N). This is achieved using a non-parametric block bootstrap approach[19,68,69], in which an additional 1000 hindcasts are created as follows:

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

Methods References

52. Morice, C. P., Kennedy, J. J., Rayner, N. A. & Jones, P. D. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.* **117**, D08101 (2012).

53. Hansen, J., Ruedy, R., Sato, M. & Lo, K. Global surface temperature change. *Rev. Geophys.* **48** (2010).

54. Karl, T. R. *et al.* Possible artifacts of data biases in the recent global surface warming hiatus. *Science* **348**, 1469–1472 (2015).

55. Schneider, U. *et al.* GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle. *Theor. Appl. Climatol.* **115**, 15–40 (2014).

56. Allan, R. J. & Ansell, T. J. A new globally complete monthly historical gridded mean sea level pressure data set (HadSLP2): 1850-2003. *J. Climate* **19**, 5816–5842 (2006).

57. Trenberth, K. E. & Shea, D. J. Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.* **33**, L12704 (2006).

58. Doblas-Reyes, F. J. *et al.* Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *Q. J. R. Meteorol. Soc.* **135**, 1538–1559 (2009).

59. Hodson, D. L. R. & Sutton, R. T. Exploring multi-model atmospheric GCM ensembles with ANOVA. *Climate Dynamics* **31**, 973–986 (2008).

60. Weisheimer, A. *et al.* How confident are predictability estimates of the winter North Atlantic Oscillation? *Q. J. R. Meteorol. Soc.* **145**, 140–159 (2019).

61. Kumar, A. & Chen, M. Causes of skill in seasonal predictions of the Arctic Oscillation. *Climate Dynamics* **51**, 2397–2411 (2018).

62. Borchert, L. F. *et al.* Decadal predictions of the probability of occurrence for warm summer temperature extremes. *Geophys. Res. Lett.* (2019).

63. Krishnamurti, T. N. *et al.* Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* **285**, 1548–1550 (1999).

64. Yun, W. T., Stefanova, L. & Krishnamurti, T. N. Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Climate* **16**, 3834–3840 (2003).

65. Kug, J.-S., Lee, J.-Y., Kang, I.-S., Wang, B. & Park, C.-K. Optimal multi-model ensemble method in seasonal prediction. *Asia-Pacific Journal of Atmospheric Sciences* **44**, 259–267 (2008).

66. Gangsto, R., Weigel, A. P., Lineger, M. A. & Appenzeller, C. Methodological aspects of the validation of decadal predictions. *Climate Res.* **55**, 181–200 (2013).

67. Smith, D., Eade, R. & Pohlmann, H. A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. *Clim. Dyn.* **41**, 3325–3338 (2013).

68. Wilks, D. S. *Statistical methods in the atmospheric sciences*, vol. 100 of *International geophysics series* (Academic Press, 2011), third edn.

69. Goddard, L. *et al.* A verification framework for interannual-to-decadal predictions experiments. *Clim. Dyn.* **40**, 245–272 (2013).

70. Doblas-Reyes, F. J. *et al.* Using EC-Earth for climate prediction research. In *ECMWF Newsletter* (ECMWF, 2018).

71. Haarsma, R. *et al.* HighResMIP versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR. Description, model performance, data handling and validation. *Geosci. Model Dev.* (submitted).

72. Counillon, F. *et al.* Flow-dependent assimilation of sea surface temperature in isopycnal coordinates with the Norwegian Climate Prediction Model. *Tellus A* **68**, 32437 (2016).

73. Wang, Y. *et al.* Optimising assimilation of hydrographic profiles into isopycnal ocean models with ensemble data assimilation. *Ocean Modelling* **114**, 33–44 (2017).

74. Kharin, V. V., Boer, G. J., Merryfield, W. J., Scinocca, J. F. & Lee, W.-S. Statistical adjustment of decadal predictions in a changing climate. *Geophys. Res. Lett.* **39**, L19705 (2012).

75. Swart, N. C. *et al.* The Canadian Earth System Model version 5 (CanESM5.0.3). *Geosci. Model Devel.* **12**, 4823–4873 (2019).

76. Sospedra-Alfonso, R. & Boer, G. J. Assessing the impact of initialization on decadal prediction skill. *Geophys. Res. Lett.* (2020).

77. Yang, X. *et al.* A predictable amo-like pattern in GFDLś fully-coupled ensemble initialization and decadal forecasting system. *J. Climate* **26**, 650–661 (2013).

78. Williams, K. D. *et al.* The Met Office Global Coupled model 3.0 and 3.1 (GC3.0 and GC3.1) configurations. *J. Adv. Model Earth Syst.* **10**, 357–380 (2018).

79. Müller, W. A. *et al.* Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. *Geophys. Res. Lett.* **39**, L22707 (2012).

80. Pohlmann, H. *et al.* Realistic Quasi-Biennial Oscillation Variability in Historical and Decadal Hindcast Simulations Using CMIP6 Forcing. *Geophys. Res. Lett.* 2019GL084878 (2019).

81. Chikamoto, Y. *et al.* An overview of decadal climate predictability in a multi-model ensemble by climate model MIROC. *Clim. Dyn.* **40**, 1201–1222 (2012).

82. Mochizuki, T. *et al.* Decadal prediction using a recent series of MIROC global climate models. *J. Meteorol. Soc. Jpn* **90**, 373–383 (2012).

Data Availability The datasets analysed during the current study are available from the CMIP data archives: https://esgf-node.llnl.gov/projects/cmip5/ and https://esgf-node.llnl.gov/projects/cmip6/. NCAR data are available from http://www.cesm.ucar.edu/projects/community-projects/DPLE/.

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

Acknowledgements DMS, AAS, NJD, LH and RE were supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra and by the European Commission Horizon 2020 EUCP project (GA 776613). FJDR, LPC, SW and RB also acknowledge the support from the EUCP project (GA 776613) and from the Ministerio de Economía y Competitividad (MINECO) as part of the CLINSA project (Grant No. CGL2017-85791-R). SW received funding from the European Union Horizon 2020 research and innovation programme under the Marie Skĺodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433 and PO from the Ramon y Cajal senior tenure programme of MINECO. The EC-Earth simulations were performed on Marenostrum 4 (hosted by the Barcelona Supercomputing Center, Spain) using Auto-Submit through computing hours provided by PRACE. WAM, HP, KM and KP were supported by the German Federal Ministry for Education and Research (BMBF) project MiKlip (grant 01LP1519A). NK, IB, FC and YW have received support from EU H2020 Blue-Action (727852), the Trond Mohn Foundation (BFS2018TMT01), the Norwegian Research Council projects INES (270061) and SFE (270733) and UNINETT Sigma2 (nn9039k, ns9039k). JR acknowledges support from NERC via NCAS and the ACSIS program (NE/N018001/1). JM, LFB and DS are supported by Blue-Action (European Union Horizon 2020 research and innovation program, Grant Number: 727852) and EUCP (European Union Horizon 2020 research and innovation programme under grant agreement no 776613) projects. The National Center for Atmospheric Research (NCAR) is a major facility sponsored by the US National Science Foundation (NSF) under Cooperative Agreement No. 1852977. NCAR contribution was partially supported by the National Oceanic and Atmospheric Administration (NOAA) Climate Program Office under Climate Variability and Predictability Program Grant NA13OAR4310138 and by the US NSF Collaborative Research EaSM2 Grant OCE-1243015. MIROC simulations were supported by MEXT through the Integrated Research Program for Advancing Climate Models (JPMXD0717935457). A.B., D.N. and P.R. were supported by the H2020 EUCP project (GA 776613).

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

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

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

Extended Data Figure 1: **Improvement of NAO-matching over variance adjustment.** (a) Time series of observed (black curve) and variance adjusted model forecast (years 2-9, red curve showing mean of the 676 member lagged ensemble and red shading showing the 5-95% confidence interval diagnosed from the forecast ensemble mean error variance) 8-year running mean December to March AMV index. (b) As (a) but for NAO-matched forecast (see Methods). (c, d) As (a, b) 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. Variance adjustment does not affect the correlation skill, but the uncertainties (red shading) capture the observations better, especially for N. Europe precipitation (compare panel c with Figure 2e). However, NAO-matching clearly improves predictions of the timing of the AMV minimum in the late 1980s and the subsequent rapid warming, and captures the observed increase in N. Europe precipitation from the 1960s to late 1980s and decrease thereafter.

Extended Data Figure 2: **Effect of NAO-matching on trends during Increasing NAO period.** Observed linear trends over hindcast start dates 1973 to 1989 inclusive for (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 standard deviation of ensemble mean 8-year means. (j, k, l) As (d, e, f) but for NAO-matched forecasts. Units are standard deviations of 8-year means per decade. The raw lagged ensemble (d, e, f) is divided by the observed standard deviation of 8-year means to show the signal relative to observed variability. NAO-matching clearly improves the cooling trend over the Labrador Sea and the warming trend over Eurasia, as well as the drying/wetting trends over southern/northern Europe.

Extended Data Figure 3: **Effect of NAO-matching on trends during decreasing NAO period.** Observed linear trends over hindcast start dates 1989 to 2005 inclusive for (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 standard deviation of ensemble mean 8-year means. (j, k, l) As (d, e, f) but for NAO-matched forecasts. Units are standard deviations of 8-year means per decade. The raw lagged ensemble (d, e, f) is divided by the observed standard deviation of 8-year means to show the signal relative to observed variability. NAO-matching improves the cooling trend over northern Eurasia, drying/wetting over northern/southern Europe, and the increasing pressure trend across most of the Arctic.

Extended Data Figure 4: **Effect of NAO-matching on skill.** Anomaly correlation skill (left panels) of 20 member NAO-matched ensemble mean, and the effect of NAO-matching (right panels), for year 2-9 boreal winter (DJFM) forecasts of (a, b) near-surface temperature, (c, d) precipitation and (e, f) mean sea level pressure (mslp). The effect of NAO-matching on skill is computed as the partial correlation between observed and forecast residuals after regressing out the lagged ensemble mean forecast[19], thereby focussing on the variability not already captured by the lagged ensemble mean. Stippling shows where correlations with observations (a, c, e) and of residuals (b, d, f) are significant (95% confidence, see Methods). Improvements from NAO-matching are consistent with the NAO-related quadrupole pattern affecting eastern North America, Greenland, western Europe, northern Africa, Eurasia, China and the Arctic. Despite the use of fewer members (20 in the NAO-matched ensemble compared to 676 in the lagged ensemble) skill is not significantly degraded in most other regions. Negative mslp skill in the Indian Ocean could be related to inconsistencies in initialisation of surface temperature and atmospheric circulation as discussed previously[19].

Extended Data Figure 5: **NAO not solely driven by AMV.** (a) Time series of observed (black curve) and variance adjusted lagged ensemble forecasts (years 2-9, red curve showing ensemble mean with shading showing 5-95% confidence interval diagnosed from the error variance) 8-year running mean December to March NAO. (b) As (a) but for AMV-matched forecasts. AMV-matching is the same procedure as NAO-matching (see Methods) except that the 20 ensemble members are selected based on AMV instead of NAO. If the NAO signal were solely driven by AMV then selecting the most skilful AMV ensemble members by AMV-matching would be expected to increase the NAO skill. However, AMV-matching clearly reduces the NAO skill (ACC reduces from 0.79, p¡0.01, to 0.37, p=0.1). In contrast, NAO-matching clearly improves the forecasts of AMV (Figure 2c and d). We therefore conclude that the NAO signal is not solely driven by AMV.

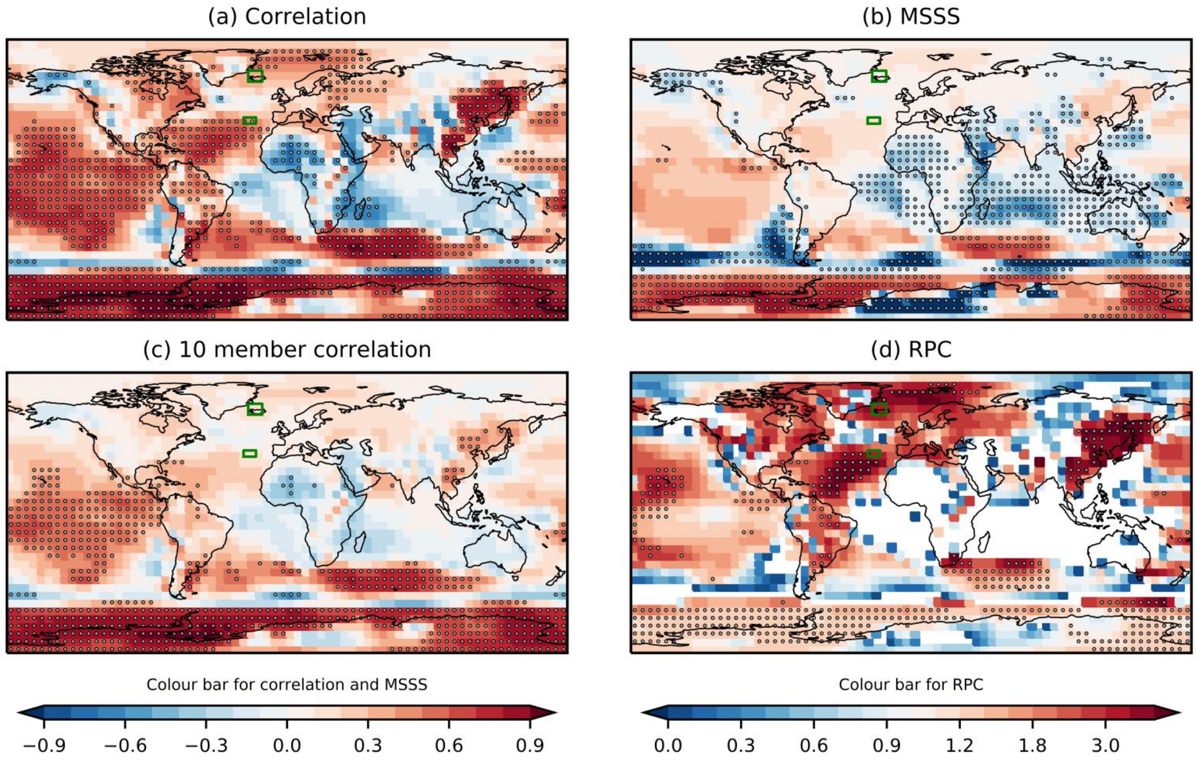

Figure 1: Decadal prediction skill for boreal winter (December to March) mean sea level pressure. Skill for year 2-9 multi-model ensemble mean forecasts measured by (a) anomaly correlation, (b) mean squared skill score (MSSS), (c) average anomaly correlation for a 10-member ensemble mean (computed over 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.

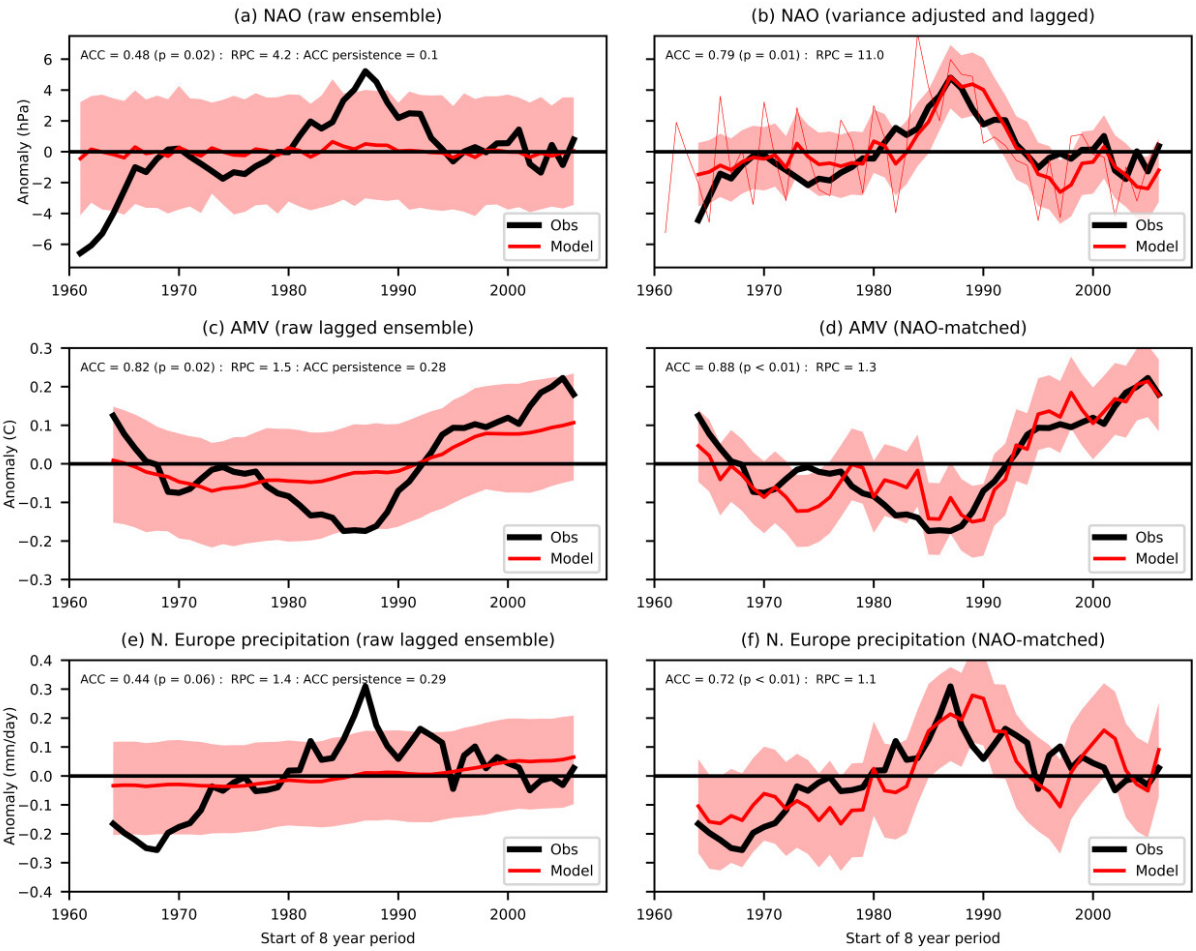

Figure 2: Underestimated signals. (a) Time series of observed (black curve) and model forecast (years 2-9, red curve showing ensemble mean of 169 members and red shading showing the 5-95% confidence interval diagnosed from the individual members) 8-year running mean December to March NAO index. (b) As (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.

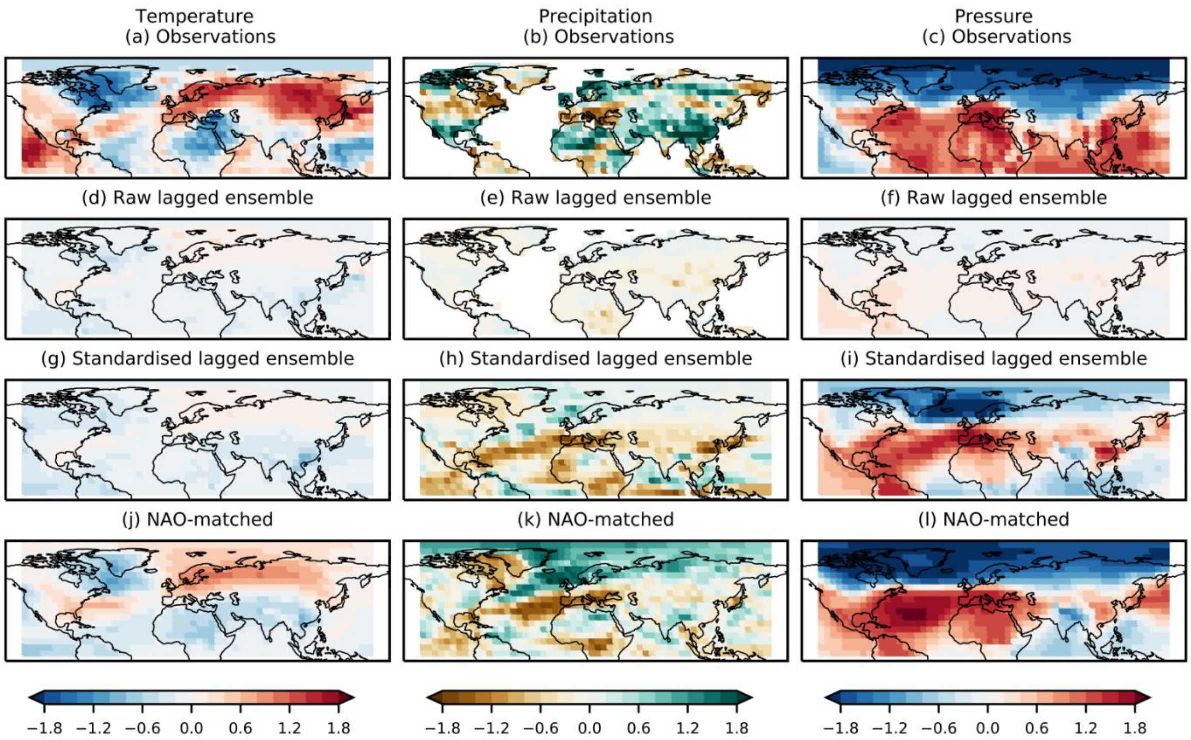

Figure 3: Decadal predictions of the extreme NAO period (1986 to 1997). Observed anomalies of (a) temperature, (b) precipitation and (c) mean sea level pressure. (d, e, f) As (a, b, c) but for raw lagged ensemble mean forecasts. (g, h, i) As (d, e, f) but standardised by the ensemble mean standard deviation. (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.