https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

Forced reorganisation of the North Atlantic SST–NAO relationship and implications for European winter predictability

Ioana Colfescu[1,2]

[1]School of Mathematics and Statistics & School of Earth and Environmental Sciences, University of St Andrews, UK

[2] National Centre for Atmospheric Science (NCAS), Leeds, UK

Correspondence: ioana.colfescu@ncas.ac.uk

AI tools were used in the preparation of this manuscript to assist with drafting and editing of text. All scientific content, analysis, results, and conclusions are the sole responsibility of the author.

Short summary

Seasonal prediction of the winter North Atlantic Oscillation (NAO) relies partly on persistent patterns of North Atlantic sea-surface temperature (SST) anomalies. Whether this SST–NAO relationship remains stable under anthropogenic climate change is unclear. Using a 40-member large ensemble together with a hierarchy of physically constrained statistical and machine-learning models, we examine how NAO predictability evolves under SSP3-7.0 forcing. The internal SST–NAO teleconnection remains broadly stable throughout the 21st century, although its subpolar component weakens. In contrast, the externally forced SST–NAO relationship changes substantially and reverses sign by late century as the North Atlantic warming pattern reorganises. Forecast skill declines most strongly during the transition period when the forced signal passes through near-zero amplitude. These results suggest that future changes in the North Atlantic mean state may alter the basis of seasonal NAO predictability even if the underlying internal teleconnection remains intact.

Abstract

The winter North Atlantic Oscillation (NAO) is a major source of seasonal climate variability across the Euro-Atlantic region, and part of its predictability arises from persistent North Atlantic sea-surface temperature (SST) anomalies. Whether this SST–NAO relationship remains stationary under anthropogenic forcing is uncertain. Here we investigate the evolution of North Atlantic SST–NAO predictability using the 40-member CANARI large ensemble under SSP3-7.0 forcing together with a hierarchy of physically constrained statistical and machine-learning models. We find that the internal component of the SST–NAO teleconnection retains a largely stable spatial structure throughout the 21st century, although the subpolar North Atlantic contribution weakens. In contrast, the externally forced SST–NAO relationship changes substantially as the North Atlantic warming pattern evolves. A three-term stationarity budget shows that forecast-skill degradation in the early 21st century is dominated by changes in the forced mean state rather than by increased internal variability. By the late century, the forced SST–NAO regression reverses sign as differential warming between the tropical and subpolar North Atlantic strengthens. A decomposition of predictability sources indicates a transition from historically interaction-driven predictability, in which forced and internal SST components jointly contribute to NAO skill, toward a regime in which the forced SST component becomes increasingly important. All model configurations exhibit a

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

comparable decline in skill under SSP3-7.0, suggesting that the degradation arises primarily from changes in the climate background state rather than from the choice of statistical architecture. ERA5 reanalysis data show an emerging positive shift in the recent winter NAO relative to the historical baseline, consistent with the early stages of the simulated transition. Together, these results suggest that future seasonal prediction systems may require periodic recalibration as the North Atlantic mean state evolves under continued anthropogenic forcing.

1. Introduction

The North Atlantic Oscillation (NAO) is the dominant pattern of atmospheric variability over the North Atlantic on weekly-to-interannual timescales (Hurrell, 1995; Hurrell et al., 2003). Its phase largely determines the position and intensity of the North Atlantic jet stream and the associated storm track. Boreal winter is the season of primary interest: the meridional temperature gradient between the tropics and the Arctic is then at its strongest, sustaining the most vigorous baroclinic activity, the most active transient eddy fluxes, and the most energetic storm track across the North Atlantic basin (Hurrell and Deser, 2010; Woollings et al., 2010). It is in this season that the NAO attains its largest amplitude and most strongly imprints on surface climate across the Euro-Atlantic sector (Hurrell, 1995; Marshall et al., 2001; Hurrell and Deser, 2010). Through its control on wind, precipitation, and temperature variability, the NAO has measurable consequences for energy demand, river flows, flood risk, and agricultural yields across Europe (Hurrell and van Loon, 1997; Visbeck et al., 2001). Skillful seasonal prediction of the winter NAO is therefore a long-standing objective of operational climate services across the Euro-Atlantic sector (Scaife et al., 2014; Kushnir et al., 2019).

A component of the winter NAO is predictable from North Atlantic sea surface temperature (SST) anomalies in the preceding summer and autumn. The canonical predictor is the North Atlantic SST tripole (Deser and Blackmon, 1993; Kushnir, 1994; Czaja and Frankignoul, 2002): cold anomalies in the subpolar gyre, warm anomalies in the subtropical gyre, and cold anomalies in the tropical North Atlantic. This pattern has been repeatedly associated with lagged NAO variability across observational and modelling studies (Visbeck et al., 2003; Sutton and Hodson, 2005; Deser et al., 2004). Through baroclinic adjustment to the diabatic heating anomalies generated by this SST pattern, the large-scale tropospheric circulation — including the position and strength of the North Atlantic jet stream — is shifted toward preferred phases of the winter NAO (Hoskins and Karoly, 1981; Peng et al., 2003; Rodwell et al., 1999; Czaja and Frankignoul, 2002; Gastineau and Frankignoul, 2015). This lag structure — autumn SST anomalies influencing the subsequent winter circulation — provides the physical basis for multi-month seasonal prediction. Coupled dynamical forecast systems that exploit this pathway achieve ensemble-mean correlations with the observed winter NAO of $r \approx 0.6$ at seasonal lead (Scaife et al., 2014) and $r \approx 0.8$ at multi-year lead following statistical post-processing (Smith et al., 2020).

These skill levels are substantial, yet a well-documented feature of coupled ensemble systems suggests they may not represent the ceiling of what is physically achievable. Ensemble forecasts tend to underestimate the true predictability of the real climate: the ensemble mean carries a weaker signal than observations actually exhibit, so the $r \approx 0.6$–0.8 values likely understate the real-world predictive potential of the SST–NAO link rather than approach it. This is the signal-to-noise paradox (Scaife and Smith, 2018; Eade et al., 2014; Smith et al., 2020; Weisheimer et al., 2024). Its origins remain debated, but its practical consequence is that even well-performing forecast systems may not be fully exploiting the ocean's influence on the winter NAO. A more fundamental question, however, concerns not the current magnitude of that skill but its long-term stability: whether the SST–

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

NAO coupling that underpins these forecasts will remain stationary as anthropogenic forcing reshapes the North Atlantic has received comparatively little attention.

The SST–NAO teleconnection does not operate in a fixed ocean state — it operates against a climatological background that is changing under anthropogenic forcing. The Atlantic Meridional Overturning Circulation (AMOC) has weakened by approximately 15% since the mid-twentieth century (Caesar et al., 2018), producing a relative cooling of the subpolar gyre — the observational fingerprint of the North Atlantic Warming Hole — while the tropical North Atlantic continues to warm under rising greenhouse gas concentrations (Bellomo et al., 2021; Menary et al., 2020; Jackson et al., 2022). The result is a progressive amplification of the tropical-to-subpolar SST gradient that projects directly onto the meridional pressure gradient the NAO modulates, and may consequently alter both the atmospheric response to SST anomalies and the background state of the jet stream through which that response is communicated to European surface climate.

This evolving ocean state sets up two physically distinct questions about the future of NAO predictability. First, does the dynamical teleconnection mechanism itself change — that is, does the atmosphere's sensitivity to a given SST anomaly pattern reorganise under forcing? Second, does the background SST state shift so far from historical conditions that a forecast system calibrated on historical data can no longer reliably interpret the SST anomalies it receives, even if the underlying sensitivity is unchanged? These two possibilities call for fundamentally different responses. A change in the mechanism would require redesigning the physical predictor. A shift in the background state, with the mechanism intact, would call for updating the climatological reference against which SST anomalies are measured.

Whether the North Atlantic SST–NAO teleconnection remains stationary under anthropogenic forcing is therefore a question with both scientific and operational significance. On the scientific side, it addresses whether a well-established physical coupling — the baroclinic response of the winter NAO to antecedent ocean heat content anomalies — survives in the altered ocean state projected for the twenty-first century. On the operational side, it determines whether the physical basis for seasonal NAO prediction will continue to be exploitable by forecast systems designed around historical observations, or whether those systems will require recalibration or reconceptualisation as the forcing evolves. Both statistical and dynamical prediction systems are calibrated primarily using historical relationships (Doblas-Reyes et al., 2013; Scaife et al., 2014), so the stationarity question has direct consequences for how those systems should be maintained and updated.

This concern is sharpened by the intrinsic signal-to-noise characteristics of the problem. The predictable, SST-forced component of the winter NAO is small relative to the background of internally generated atmospheric variability — predictable variance accounts for only a few percent of total NAO variance in coupled model ensembles. As a consequence, even modest changes in the amplitude or spatial structure of the forced SST signal can substantially alter forecast skill: a contraction of the forced regression toward zero is sufficient to suppress predictability, while a sign reversal produces an entirely new predictability regime. Distinguishing between these possibilities — whether future skill changes reflect increased atmospheric noise, weakening of the forced SST amplitude, or a structural reorganisation of the SST–NAO relationship — is essential for understanding whether any degradation can be addressed through recalibration or requires a more fundamental revision of the prediction approach.

Answering this question requires separating two components of the SST–NAO relationship that are conflated in any single observation or simulation: the internally driven, dynamical component — the baroclinic mechanism

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

excited by internal variability in ocean heat content — and the externally forced component, reflecting the NAO tendency driven by the anthropogenic and natural forcing agents common to all ensemble members. If the internal mechanism is preserved but the forced component changes sign or amplitude, the appropriate response differs fundamentally from the case in which both change together.

We address these questions using the 40-member CANARI large ensemble (HadGEM3-GC3.1; Walters et al., 2019), which provides the large sample sizes required to cleanly separate the externally forced and internal-variability components of the SST and NAO at each timestep. Our primary diagnostic tool is a three-term stationarity budget that partitions the change in seasonal forecast skill between the historical period and each future evaluation window into contributions from the forced SST–NAO regression amplitude, the internal variability amplitude, and the spatial reorganisation of the forced regression. This budget closes additively, providing a complete attribution of skill change without residual.

To apply the budget, we use a hierarchy of diagnostic prediction models — spanning from a simple linear regression to a physics-anchored probabilistic network — not as competing forecast systems but as a consistency check: if all configurations degrade in the same way under future forcing, that degradation reflects a property of the SST–NAO relationship rather than any modelling choice. A complementary predictability-source analysis separates the contributions to skill from the forced and internal-variability components of the SST field, revealing how the origin of seasonal predictability shifts across the century.

Statistical and machine-learning predictors fitted on historical SST–NAO relationships offer a structural advantage for this diagnostic purpose: because their regression weights are explicit and observable, their stability across different forcing conditions provides a direct empirical test of teleconnection stationarity. This role is complementary to, rather than in competition with, dynamical coupled models, in which the SST–NAO relationship is realised implicitly through resolved physical processes.

The paper is organised as follows. Section 2 describes the data, diagnostic models, and budget framework. Section 3 presents results: prediction model validation (§3.1), internal teleconnection robustness (§3.2), the stationarity budget and sign reversal (§3.3), the predictability-source transition (§3.4), spatial attribution (§3.5), observational constraints (§3.6), and model-configuration consistency (§3.7). Section 4 discusses the physical interpretation, implications, and limitations.

2. Data and Methods

2.1 CANARI large ensemble and observational data

The primary dataset is the CANARI large ensemble, a 40-member ensemble of historical (HIST; 1950–2014) and future (SSP3-7.0; 2015–2099) simulations carried out with the HadGEM3-GC3.1 fully coupled climate model (Walters et al., 2019). CANARI was designed specifically to investigate internally generated variability and predictability under evolving anthropogenic forcing through the use of a large number of ensemble members initialised from slightly perturbed atmospheric conditions. The use of a large ensemble framework is particularly important for the present study because it permits the externally forced response, estimated as the ensemble mean, to be separated from internally generated climate variability of individual ensemble members. This separation is essential for diagnosing whether changes in seasonal predictability arise from modifications to the underlying internal teleconnection or from changes in the evolving forced climate background state.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

The winter NAO index is computed following Smith et al. (2020) as the monthly mean sea-level pressure difference (in Pa) between the Azores (28°–20°W, 36°–40°N) and Iceland (25°–16°W, 63°–70°N) action centres. The focus throughout is on boreal winter (December–March; DJFM). Predictors for the diagnostic models are North Atlantic SSTs over the domain 20°–70°N, 80°W–10°E, evaluated at four monthly lags prior to the DJFM season. SST and NAO are expressed as anomalies relative to a 1950–1980 climatology and standardised by the training-period statistics; the forced trend within each analysis window is retained. Sensitivity tests using detrended anomalies produce qualitatively similar large-scale behaviour as described in Section 2.2, although the late-century forced subpolar amplitude is reduced (Fig. 2c–d).

To separate externally forced from internally generated variability, each ensemble member is decomposed into an ensemble-mean component and a residual component after subtracting the ensemble mean. The ensemble-mean term represents the externally forced response, and the residual represents the internally generated anomaly after removal of the ensemble-mean forced signal. This decomposition is applied consistently to both SST and NAO fields. Ensemble-mean quantities are interpreted as approximations to the forced climate response under SSP3-7.0 forcing, whereas deviations from the ensemble mean are interpreted as internally generated variability arising from atmosphere–ocean interactions and chaotic atmospheric dynamics. The validity of this decomposition relies on the assumption that internally generated variability is uncorrelated across ensemble members and therefore averages toward zero in the ensemble mean. While residual contamination from internal variability cannot be completely excluded, previous large-ensemble studies have shown that ensemble means constructed from approximately 30–40 members provide a robust estimate of the forced North Atlantic climate response on multidecadal timescales.

In the sections below, the term “forced relationship” refers to regressions computed using ensemble-mean quantities, whereas the term “internal relationship” refers to regressions computed from ensemble anomalies after removal of the forced signal. The decomposition additionally permits the covariance between forced and internal components to be evaluated explicitly. This is important because predictability may arise not only from the internally generated SST signal itself, but also from interactions between internally generated anomalies and the evolving forced mean state.

Throughout this work, σ_{signal} denotes the standard deviation over winters of the ensemble-mean (forced) NAO time series — a measure of the externally forced NAO amplitude. σ_{noise} is the root-mean-square across ensemble members of each member's internal NAO standard deviation (NAO minus ensemble mean). The ratio $SNR \equiv \sigma_{signal} / \sigma_{noise}$ indicates the fraction of total NAO variance that is predictable from external forcing. Over 1950–2014 this gives $\sigma_{signal} \approx 133$ Pa, $\sigma_{noise} \approx 782$ Pa, and $SNR \approx 0.17$, so predictable variance is roughly 3% of total NAO variance.

To evaluate the consistency with observations, we used ERA5 (Hersbach et al., 2020) for the NAO index and HadISST (Rayner et al., 2003) for SST, both over 1950–2024. Note that observations are used here primarily as a qualitative benchmark for the historical SST–NAO relationship and recent circulation trends rather than as a strict validation target because the observational period remains short relative to multidecadal North Atlantic variability.

Forecast skill is evaluated separately over historical (1950–2014), near-future (2015–2039), mid-century (2040–2064), and late-century (2065–2099) periods in order to diagnose the evolution of predictability across progressively warmer climate states.

Before applying any diagnostic models, we verified that the CANARI ensemble reproduces the observed lagged SST–NAO relationship in the historical period. Lead/lag regression maps of the winter NAO index onto

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

preceding monthly North Atlantic SST anomalies were computed for CANARI (ensemble-mean of member-wise regressions, 1950–2014) and compared with the equivalent ERA5–HadISST observational estimate over the same period. The CANARI historical regression captures the canonical tripole structure — subpolar cooling, subtropical warming, and tropical cooling ahead of positive NAO winters — with amplitude and sign consistent with the observational reference (not shown; Czaja and Frankignoul, 2002; Rodwell et al., 1999). Both amplitude and lag of the regression are broadly consistent with the observational reference. This agreement confirms that the CANARI ensemble is a suitable basis for the stationarity analysis that follows. The same lead/lag regression is shown for the SSP3-7.0 period (not shown), illustrating that the spatial structure of the internal teleconnection is preserved under future forcing — consistent with the pattern-correlation analysis reported in Section 3.2 — while the amplitude in the subpolar node declines moderately. Taken together, these comparisons establish that changes diagnosed in §3 reflect genuine evolution of the SST–NAO relationship under forcing, rather than a pre-existing deficiency in how the model represents the teleconnection.

2.2 Diagnostic prediction model hierarchy

The model hierarchy is designed primarily as a diagnostic framework rather than as a competition between prediction architectures. Each model level introduces one additional degree of flexibility into the SST–NAO relationship, allowing us to assess whether the conclusions depend on model complexity or instead reflect properties of the underlying climate signal itself.

The simplest configuration (L0) is a ridge regression fitted to the leading principal components of North Atlantic SST variability. This configuration tests whether a linear SST signal is sufficient to recover seasonal NAO predictability. The second configuration (L1) replaces the linear regression with a convolutional neural network applied directly to the gridded SST field at multiple monthly lags, allowing nonlinear and lag-dependent relationships to emerge. The third configuration (L2) retains the same convolutional structure but predicts a Gaussian distribution rather than a single deterministic NAO value, thereby representing forecast uncertainty explicitly. The final configuration (L3) further allows the predictive distribution to be non-Gaussian and tests whether constraining the learned SST representation to the known SST–NAO tripole pattern affects the result.

Essentially, each tier adds one capability to the SST–NAO mapping:

- L0 (linear regression): Is there a detectable linear SST–NAO signal under low signal-to-noise conditions?
- L1 (convolutional network with lags): Does allowing nonlinearity and lag-dependent weighting improve skill?
- L2 (probabilistic convolutional network): Does representing forecast uncertainty — rather than a single best estimate — improve the skill estimate?
- L3 (physics-anchored network): Does constraining the learned representation to the known SST–NAO tripole pattern help, at the cost of flexibility?

The purpose of the hierarchy is not to identify a single optimal predictor, but to test whether changes in forecast skill are robust across fundamentally different statistical representations. If all model levels exhibit similar behaviour under climate change, then the resulting predictability changes are unlikely to arise from any individual architecture and instead more likely reflect changes in the underlying SST–NAO relationship.

All configurations are trained on CANARI members 1–32 (1950–1999) and applied without retraining, fine-tuning or recalibration to a held-out test set (member33-40) over 2000–2014 and to each of the SSP3-7.0 windows

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

(2015–2039, 2040–2064, 2065–2099). SST inputs in the SSP3-7.0 period are normalised using HIST training statistics, so the only quantity changing between training and evaluation is the climate state itself. This experimental design mirrors the practical problem faced by seasonal forecast systems trained under the historical climate and subsequently applied in a changing background state. Skill is reported as the Pearson correlation between predicted and observed NAO across all member-winter pairs in each window. Full technical specifications — loss function, training protocol, architecture details — are given in Appendix A. A sensitivity experiment repeating all analyses after subtracting the ensemble-mean linear SST trend confirms that the results are robust to detrending (Supplementary Fig. S1).

Three aspects of the low signal-to-noise environment require specific design choices, described briefly here and specified fully in Appendix A. First, the training target: because the ensemble-mean NAO has near-zero variance at SNR ≈ 0.17, it cannot serve as a meaningful training signal. We instead use, at each timestep, the ensemble member whose NAO departs most strongly from the ensemble mean — a choice that preserves the direction of the forced signal while providing sufficient variance for training, with an amplitude correction reanchoring the predicted variability to the observed NAO magnitude. Second, the loss function is supplemented with a penalty on the predicted variance, preventing the models from collapsing toward a climatological mean rather than responding to the SST input. Third, for the most constrained configurations (L2, L3), a fixed projection layer initialised from the observed SST–NAO regression tripole constrains the learned SST representation to the known physical teleconnection pattern, guarding against spurious correlations that can arise at low signal-to-noise (Reichstein et al., 2019; Karniadakis et al., 2021). L0 and L1 are left unconstrained, so the hierarchy spans the full range from unconstrained to physics-anchored. Full specifications are in Appendix A.

2.3 Three-term stationarity budget and predictability-source decomposition

To diagnose changes in predictability under climate change, we construct a three-term stationarity budget that separates forecast-skill changes between the historical period and each SSP3-7.0 windows intro additive contributions associated with:

- changes in forced signal amplitude,
- changes in signal-to-noise ratio,
- changes in the spatial structure of the SST–NAO relationship.

The framework is designed to distinguish between reduced predictability arising from increased internally generated atmospheric variability and reduced predictability arising from structural evolution of the SST-forced circulation response under anthropogenic forcing. Forecast degradation can occur through several physically distinct pathways. First, the amplitude of the predictable SST signal may weaken relative to historical conditions. Second, internally generated atmospheric variability may increase, thereby reducing the effective signal-to-noise ratio available to seasonal prediction systems. Third, the SST–NAO teleconnection itself may evolve spatially, such that historical predictor structures become progressively less representative of future climate states even if the underlying teleconnection mechanism remains partially intact.

Formally, this is written as: $\Delta r_{total} = \Delta r_{signal} + \Delta r_{noise} + \Delta r_{pattern}$. They are built from the forced-regression correlation r — the Pearson correlation between the ensemble-mean SST time series $\langle SST(t) \rangle$ and the ensemble-mean NAO time series $\langle NAO(t) \rangle$. The ensemble mean at each timestep is used as the operational

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

estimate of the externally forced response (Hawkins and Sutton, 2009; Deser et al., 2012); ensemble-averaged member-wise regressions are not used because they carry an internal-variability bias of order var(internal)/var(total) that prevents budget closure at CANARI's low SNR ≈ 0.17. The forced regression slope β = cov(⟨SST⟩, ⟨NAO⟩) / var(⟨SST⟩) [Pa K^{-1}] is a diagnostic derived from the same ensemble-mean fields. Two variance quantities defined in §2.1 enter the noise term: σ_signal, the standard deviation over winters of the ensemble-mean NAO time series (forced NAO amplitude); and σ_noise, the root-mean-square across the 40 members of each member's standard deviation over winters of the internal NAO (NAO minus ensemble mean; internal variability amplitude).

The total skill change is Δr_total = r_HIST − r(w), where r(w) is the forced-regression correlation in window w computed on the raw SST anomaly fields (anomalies relative to the fixed 1950–1980 HIST climatology, retaining the forced SST trend). The three terms decompose this total as follows. Δr_signal = r_HIST − r_signal(w), where r_signal(w) is r(w) recomputed after subtracting the per-grid-point ensemble-mean SST at each timestep, removing the forced trend from the anomaly field so that only the change in forced SST amplitude contributes. Δr_noise = r_HIST − r_noise(w), where r_noise(w) is the forced-regression correlation recomputed after rescaling the SSP3-7.0 NAO variance to the HIST σ_signal and σ_noise values, isolating the contribution from the change in the ratio σ_signal / σ_noise while holding the regression pattern fixed; a lower SNR reduces the forced-regression correlation even when the physical pattern is unchanged. Δr_pattern captures the reorganisation of the spatial structure and sign of β that remains after amplitude and noise contributions have been removed. All three budget terms are each computed independently from their own estimators and are not derived sequentially from one another, and they are verified to close ($|\Delta r\_total - (\Delta r\_signal + \Delta r\_noise + \Delta r\_pattern)| < 10^{-3}$ in all windows; Appendix B).

Changes in teleconnection structure are quantified using centred spatial pattern correlations between historical and future SST–NAO regression maps over the North Atlantic domain. Pattern correlations are computed separately for the full-field, forced, and internal components in order to determine whether structural evolution arises primarily from externally forced changes or from modification of the internally generated teleconnection itself.

The predictability-source decomposition identifies where skill originates within the SST field. Three regression experiments, identical except for their SST input, are run using the L0 configuration: one using only the ensemble-mean (forced) SST, one using only the per-member departure from the ensemble mean (internal-variability SST), and one using the full per-member SST. The interaction term — combined skill minus the sum of the forced and internal-variability contributions — captures skill that arises only from the joint structure of the two components, and cannot be recovered from either alone. L0 is used for this decomposition because its linearity keeps the terms analytically separable.

The decomposition is intended primarily as a physically interpretable diagnostic framework rather than as an exact variance partition. Interactions between forced and internal variability may generate covariance terms that are not fully separable within a linear decomposition framework. Nevertheless, the stationarity budget provides a useful approximation for identifying the dominant contributors to changing forecast skill under anthropogenic forcing.

Uncertainty ranges are estimated using bootstrap resampling across ensemble members together with temporal block resampling across winters in order to account for both ensemble spread and temporal autocorrelation.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

Statistical significance of differences between historical and future windows is evaluated using two-sided bootstrap confidence intervals.

3. Results

3.1 Prediction model validation

Before examining how skill evolves under future forcing, we establish that the diagnostic model hierarchy reproduces the known SST–NAO relationship during the historical period. Figure 1 summarises model performance on the held-out test set (ensemble members 33–40, 2000–2014).

The simplest configuration, L0 (linear regression), achieves the highest time-series skill in the historical period ($r = 0.19$, 95% CI: 0.11–0.27; Fig. 1a). Skill decreases monotonically with model complexity across L1–L3. This inverse relationship between complexity and skill is a direct consequence of the low ratio of predictable variance to internal variability in CANARI (signal-to-noise ratio SNR ≈ 0.17): with the predictable signal accounting for only about 3% of total NAO variance, more flexible models tend to fit the unpredictable internal variability in the training data rather than the small physical signal. An analogous tendency is well documented in the coupled seasonal forecast literature (Scaife and Smith, 2018; Weisheimer et al., 2024). That the hierarchy reproduces this known behaviour gives confidence that the models are responding to the physical SST–NAO teleconnection rather than to spurious features of the training data.

A complementary picture emerges when the spatial structure of the learned SST predictor is compared with the observed SST–NAO tripole regression. L0 achieves its time-series skill through a combination of SST variance modes that does not closely resemble the physical tripole pattern (spatial pattern correlation $r = -0.05$). The more constrained configuration L2, by contrast, recovers the canonical tripole — subpolar cooling, subtropical warming, and tropical cooling ahead of positive NAO winters — with a pattern correlation of $r = +0.58$ against the ensemble-estimated regression (Fig. 1c,d). This recovery comes at a small cost to time-series skill ($\Delta r \approx 0.04$ relative to L0). The two configurations are therefore complementary: L0 provides the most reliable estimate of temporal skill and is used as the baseline for the stationarity budget in §3.3, while L2 provides spatially interpretable regression maps that can be compared with the known physical teleconnection pattern.

The frequency characteristics of the L0 predictions also match the observed NAO (Fig. 1b): the power spectrum of the predicted NAO closely follows that of the target NAO across all resolved frequencies, and both lie substantially above a first-order autoregressive baseline. This confirms that the model is exploiting interannual-to-decadal variability in the SST field rather than simply fitting the year-to-year persistence of the NAO itself. The preferential weighting of longer-lag SST information by the convolutional configurations is consistent with the well-established role of upper-ocean heat content memory in the subpolar gyre as a multi-month precursor of the winter NAO (Czaja and Frankignoul, 2002; Gastineau and Frankignoul, 2015). Together, these validation results confirm that the hierarchy captures the known SST–NAO relationship in the historical period and can therefore serve as a reliable diagnostic for the stationarity analysis that follows.

3.2 The internal teleconnection is preserved, the forced shift is already detectable

The internal-variability component of the SST–NAO teleconnection — the dynamical pathway by which ocean heat content anomalies modulate the tropospheric circulation — remains spatially coherent throughout the twenty-first century under SSP3-7.0 forcing (Fig. 2a–d). Pattern correlations between the internal teleconnection maps

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

and the historical reference decline monotonically but gradually, from r = 0.96 in 2015–2039 to r = 0.87 in 2065–2099 (overall r = 0.916 across the full SSP3-7.0 period; Fig. 2e). The attenuation is geographically concentrated in the subpolar node (50–65°N), consistent with the progressive weakening of the subpolar gyre under AMOC decline (Caesar et al., 2018; Menary et al., 2020), which reduces the ocean-to-atmosphere heat flux available to force the troposphere in that region. The tropical and subtropical nodes of the tripole are comparatively stable. No abrupt spatial reorganisation of the internal teleconnection is apparent at any point in the century. The physical coupling mechanism is preserved.

This preservation has an important implication: any future skill loss in historically calibrated seasonal forecast systems cannot be attributed to a breakdown of the physical ocean-to-atmosphere pathway. What has changed — as the following sections demonstrate — is the climatological SST background against which that pathway operates.

ERA5 reanalysis indicates that changes in the forced mean state are already underway. The DJFM NAO index shows a positive shift of +527 Pa over 2015–2024 relative to the 1950–2014 mean (95% CI: 84–1002 Pa, excluding zero; Fig. 6b,c). This anomaly lies outside the CANARI 95% ensemble range (Fig. 6d), consistent with the known tendency for coupled models to underestimate the predictable forced signal (Smith et al., 2020; Weisheimer et al., 2024). The direction of the ERA5 shift is consistent with the early-SSP3-7.0 forced response diagnosed in §3.2, providing initial observational support for the forced transition.

3.3 Three-term stationarity budget and forced response sign reversal

Despite the preserved internal teleconnection, seasonal forecast skill declines comparably across all SSP3-7.0 windows and across all four prediction configurations (§3.6; Fig. 3). This decline across different architectures emphasises the presence of non-stationarity as a property of the climate system rather than of any individual statistical method. The three-term budget identifies the physical origin of this decline.

The internal NAO variability amplitude is stationary throughout: σ_noise declines by less than 2% across all windows (782 to 773 Pa) (Fig. S2). Changes in internal variability play no role in the skill evolution. The budget is therefore governed by the signal and pattern terms, and their relative contributions change markedly across the century.

In 2015–2039, approximately 91% of the total skill decline is associated with the signal term (Fig. 3f). During this window the forced regression slope passes through zero ($\beta \approx -70$ Pa K^{-1}). Physically, this marks the transition between two forced regimes. Under the broadly uniform North Atlantic warming of the historical period, the subtropical-to-subpolar pressure gradient progressively weakens, giving a negative forced NAO tendency — the externally forced component of SST anomalies is associated with a tendency toward negative NAO winters. As uniform warming gives way to AMOC-driven differential warming, this forced tendency contracts. The physical mechanism linking internal SST anomalies to NAO variability is functioning as before; what has changed is that the forced component of SST, which historically modulated this link, has temporarily lost its coherent sign.

By 2065–2099 the transition is complete and a qualitatively different forced regime is established (Fig. 3c,d). AMOC weakening has driven differential warming in which the tropical North Atlantic warms considerably faster than the subpolar gyre (Caesar et al., 2018; Menary et al., 2020; Bellomo et al., 2021). This pattern strengthens the meridional pressure gradient and reverses the sign of the forced NAO tendency, so that the externally forced SST anomalies are now associated with positive NAO winters (β: −263 to +306 Pa K^{-1}). The forced signal re-amplifies,

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

and the pattern term is now co-equal with the signal term (59% signal, 45% pattern; Fig. 3e,f), reflecting the spatial reorganisation accompanying the sign reversal. Throughout this evolution, the noise term remains negligible, confirming that the changing predictability regime is driven entirely by the forced SST component.

3.4 Predictability-source transition: from interaction-driven to forced-driven

The predictability-source decomposition reveals how the physical origin of seasonal NAO skill shifts across the century (Fig. 4). In the historical period, neither the forced SST anomaly alone nor the internal-variability SST alone yields meaningful skill ($r \approx 0$ in each case), yet the combined SST field achieves $r = 0.16$. The interaction term accounts for essentially all of this: $r_{interaction} \approx +0.16$. Historical predictability arises not from either SST component acting independently, but from their joint structure — specifically, the way the forced background state modulates the conditions under which internal SST anomalies excite the NAO. This is physically consistent with the small historical forced amplitude (SNR ≈ 0.17): the forced signal is too weak to drive predictability on its own, but it shapes the background ocean state in a way that determines how the NAO responds to internal variability.

By 2065–2099 this source composition has fundamentally changed. The forced SST component alone yields $r = +0.16$ (95% CI: 0.09–0.23), accounting for most of the late-century skill; the internal-variability component contributes $r = +0.04$ and the interaction term is substantially reduced ($r_{interaction} = +0.09$ vs. $+0.16$ in HIST; Fig. 4b). The emergence of direct forced predictability reflects the re-amplified, sign-reversed forced regression: the forced SST component is now large enough to drive NAO predictability on its own, without requiring the modulation of internal variability. The total apparent skill is similar in both periods ($r \approx 0.16$), but it comes from entirely different physical sources. A forecast system calibrated on the historical regime — in which predictability required the joint structure of forced and internal SST — is not designed to exploit this new forced predictability, though it may capture it incidentally.

3.5 Spatial attribution of predictability

Regional masking and network sensitivity analysis reveal how the spatial footprint of NAO predictability shifts across the forcing period (Fig. 5). In the historical period, the Gulf Stream extension and North Atlantic Current (35°–50°N, 75°–30°W) contribute most strongly to the skill (reduction in skill of $r = +0.023$ when this region is masked), consistent with the baroclinic forcing of the NAO by the western boundary current SST front (Czaja and Frankignoul, 2002).

In the 2015–2039 transition window this relationship inverts: masking the Gulf Stream region marginally improves skill ($\Delta r = -0.016$), indicating that Gulf Stream SST variability is projecting more onto prediction noise than onto the near-zero forced signal during this transitional period. This is the spatial expression of the 91% signal-term dominance: the historical predictability pathway has been disrupted, and the new forced pathway is not yet established.

By 2065–2099 the footprint of predictability broadens substantially. The subpolar gyre, Gulf Stream, and mid-latitude Atlantic (30°–45°N, 50°–10°W) each contribute comparably ($\Delta r \approx +0.012$–0.015 when masked), and the eastern Atlantic and European margin (40°–65°N, 30°W–10°E) shows a statistically detectable increase in sensitivity. This broadening directly reflects the geographic structure of the AMOC-mediated differential warming — subpolar weakening on one node, tropical and subtropical strengthening on the other — which reorganises the full SST tripole in amplitude and sign. The increasing predictive weight of SSTs immediately west of the United

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

Kingdom, in the sector most directly coupled to NAO-driven precipitation and temperature anomalies over the British Isles, is of direct relevance to European climate services. The tropical North Atlantic contributes negligibly in all periods ($|\Delta r| \leq 0.003$), consistent with the NAO retaining its extratropical character throughout.

3.6 Observational constraints from ERA5

Figure 6a shows the ERA5 DJFM NAO index from 1950 to 2024 alongside the CANARI 5–95% ensemble envelope. ERA5 indicates that changes in the forced mean state are already underway: the DJFM NAO tracks within the CANARI ensemble envelope across the historical period but lies persistently above the ensemble mean during 2015–2024, with several recent winters approaching the upper bound of the simulated range (Fig. 6a). The ERA5 positive NAO shift of +527 Pa over 2015–2024 relative to 1950-2014 (95% CI: 84–1002 Pa; Fig. 6b,c) is consistent with the CANARI early-SSP3-7.0 forced response in two respects. The direction of the shift agrees with the budget: as the forced regression transitions from negative toward positive, passing through zero, a positive shift in the mean NAO is the expected early signature. Also, the ERA5 amplitude exceeds that of any individual CANARI member over the same period (Fig. 6d), which is consistent with the tendency for models to underestimate the magnitude of the predictable forced signal rather than indicating an error in the direction of the projected change. This is the observational counterpart of the signal-to-noise paradox: predictable variance in the real world exceeds predictable variance in the model.

The 10-year ERA5 record is at the lower limit for formal attribution and should be revisited as observations accumulate. Whether the ERA5 NAO trajectory will continue to follow the CANARI ensemble mean direction, or diverge further consistent with model underestimation of the forced signal, is the empirical test the next decade of observations will provide. The budget predicts near-zero diagnostic skill through the 2015–2039 transition window, followed by partial recovery as the sign-reversed forced regime establishes itself.

3.7 Consistency of skill decline across prediction configurations

All four prediction configurations (L0–L3) show comparable skill declines across the SSP3-7.0 windows, with no statistically distinguishable differences between configurations (pairwise test $p > 0.49$; largest within-window difference: $\Delta r = 0.037$ between L0 and L3; Appendix B). Within the historical period, skill is highest for the simplest configuration (L0; $r = 0.19$, 95% CI: 0.11–0.27) and declines monotonically with complexity — consistent with additional model parameters capturing internal variability rather than the weak forced signal. The more constrained configurations recover the spatial SST–NAO tripole pattern more faithfully (L2 pattern correlation: $r = +0.58$), at a small cost to time-series skill. These results are consistent with the comparable declines across configurations arising from the properties of the changing SST–NAO relationship itself, not from configuration-specific limitations.

4. Discussion and conclusions

The results presented here suggest a physically coherent picture of how European winter seasonal predictability may evolve under continued anthropogenic forcing. The central finding is a distinction between two components of the SST–NAO relationship that respond differently to forcing: the internal, dynamically driven teleconnection is preserved in spatial structure throughout the twenty-first century, while the forced component undergoes a sign

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

reversal after mid-century. This distinction matters because it identifies the source of future skill loss as a shift in the ocean mean state, not a breakdown of the physical coupling mechanism.

The physical sequence is as follows. During the historical period, greenhouse gas forcing drives broadly uniform warming across the North Atlantic basin. This progressively weakens the subtropical-to-subpolar meridional pressure gradient, producing a negative forced NAO tendency — one that is too weak (SNR ≈ 0.17) to drive predictability independently, but sufficient to modulate the background ocean state through which internal SST variability influences the NAO. As AMOC weakening accelerates the differential warming between the tropical and subpolar North Atlantic (Caesar et al., 2018; Menary et al., 2020; Bellomo et al., 2021), this forced tendency contracts toward zero in 2015–2039 and reverses sign by late century. The result is a qualitatively new forced predictability regime in which the externally forced SST component, now large and positively signed, drives NAO predictability independently of internal variability. The physical mechanism is intact throughout; it is the forced ocean context in which it operates that has changed.

This framing has direct implications for how seasonal forecast systems should respond. Systems calibrated on the historical period will encounter two distinct challenges. In the near term (2015–2039), the forced SST mean state is transitioning through a low-amplitude regime in which the historical calibration no longer accurately reflects the current predictability environment. ERA5 already provides observational evidence of this transition: the +527 Pa NAO shift over 2015–2024 lies outside the full model ensemble range, consistent with an observed forced shift that is larger than any individual CANARI member projects. By late century, a sign-reversed forced regime offers new predictability, but one that a historically calibrated system was not designed to exploit. In both cases, updating the climatological SST reference against which anomalies are computed — rather than reconceptualising the physical predictor — is likely the most effective operational response.

The stationarity budget developed here is not specific to this particular model or teleconnection. Applied alongside skill assessments in any seasonal forecast setting, it can identify whether an observed skill change reflects a structural shift in the forced amplitude or regression pattern — requiring climatological recalibration — or a change in the internal variability level — which would not be addressable through recalibration alone. The consistency test across prediction configurations (§3.7) confirms that it reflects a property of the underlying climate dynamics rather than any statistical modelling choice#.

Several caveats qualify how firmly these results generalise. The budget split and the timing of the sign reversal are specific to CANARI, which is based on HadGEM3-GC3.1 with an effective climate sensitivity of 5.5 K (Andrews et al., 2019) — placing it among the higher-sensitivity CMIP6 models (cf. CESM2 at 5.3 K, Gettelman et al., 2019; CanESM5 at 5.6 K, Swart et al., 2019). A higher climate sensitivity drives a more pronounced and potentially earlier shift in the background SST state, so the timing of the sign reversal within CANARI may represent an upper bound relative to lower-sensitivity models. Models with a weaker or slower AMOC decline would project a later crossing, with direct consequences for when operational forecast systems enter the transition window. Testing the three-term budget across additional large ensembles (e.g., CESM2-LE, CanESM5-LE) is the natural multi-model extension of this work. The observational evidence rests on a 10-year ERA5 record at the lower limit for formal attribution, and should be revisited as the post-2024 record accumulates. The budget also assumes that the forced and internal SST contributions are approximately additive — an assumption verified empirically at the low signal-to-noise ratios characteristic of CANARI, but not guaranteed under higher emissions where forced-internal coupling may strengthen.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

Three open questions follow. Whether the timing of the forced-regression sign change is robust across the CMIP6 ensemble — and therefore when operational forecast systems can expect to enter the near-zero-skill transition window — remains to be established. Whether an analogous interaction-to-forced-driven predictability transition occurs along the ENSO–NAO pathway, where future sign changes have been reported (Geng et al., 2024), is a natural extension. And whether a forecast system explicitly recalibrated against the post-2014 SST background could already begin exploiting the excess forced NAO signal visible in ERA5 is the most immediately operationally relevant question raised by this analysis.

In summary, the internal SST–NAO teleconnection that underpins European winter seasonal predictability appears likely to remain physically intact through the twenty-first century. What is not stable is the forced ocean mean state against which it operates. As AMOC-driven differential warming progressively reverses the sign of the forced SST–NAO regression, a new forced predictability regime is expected to emerge by late century. ERA5 already shows early signs of this transition. The results suggest that recognising and responding to this shift in the forced background state — through periodic recalibration rather than structural redesign — may be the most effective way to maintain the skill of European winter seasonal forecasts in a changing climate.

Code and Data Availability

CANARI output is made available through the CEDA/JASMIN repository (https://catalogue.ceda.ac.uk/uuid/0c80814cd7ec4f15a89629c582968b74/). ERA5 data is available at https://cds.climate.copernicus.eu/. The HadISST dataset is available at https://www.metoffice.gov.uk/hadobs/hadisst/. Code: [GitHub DOI].

Acknowledgements

The author acknowledges JASMIN (https://www.jasmin.ac.uk), operated by STFC RAL Space on behalf of Natural Environment Research Council (NERC). CANARI simulations were produced by the UK Met Office. This work was supported by NCAS through CANARI. The author acknowledges the use of generative AI tools to assist with some of the figures' generation and to polish the English-language clarity of earlier/initial manuscript drafts; all scientific content, analyses, and conclusions are the author's own, with full responsibility for the final text.

Author Contributions

I.C.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing — original draft, Writing — review and editing.

Competing Interests

The author declares no competing interests.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

References

Andrews, T., Andrews, M. B., Bodas-Salcedo, A., Jones, G. S., Kuhlbrodt, T., Manners, J., Menary, M. B., Ridley, J., Ringer, M. A., Sellar, A. A., Senior, C. A., and Tang, Y.: Forcings, feedbacks, and climate sensitivity in HadGEM3-GC3.1 and UKESM1, J. Adv. Model. Earth Syst., 11, 4377–4394, https://doi.org/10.1029/2019MS001866, 2019.

Bellomo, K., Angeloni, M., Corti, S., and von Hardenberg, J.: Future climate change shaped by inter-model differences in Atlantic meridional overturning circulation response, Nat. Commun., 12, 3659, https://doi.org/10.1038/s41467-021-24015-w, 2021.

Caesar, L., Rahmstorf, S., Robinson, A., Feulner, G., and Saba, V.: Observed fingerprint of a weakening Atlantic Ocean overturning circulation, Nature, 556, 191–196, https://doi.org/10.1038/s41586-018-0006-5, 2018.

Czaja, A. and Frankignoul, C.: Observed impact of Atlantic SST anomalies on the North Atlantic Oscillation, J. Clim., 15, 606–623, https://doi.org/10.1175/1520-0442(2002)015<0606:OIOASA>2.0.CO;2, 2002.

Deser, C. and Blackmon, M. L.: Surface climate variations over the North Atlantic Ocean during winter: 1900–1989, J. Clim., 6, 1743–1753, https://doi.org/10.1175/1520-0442(1993)006<1743:SCVOTN>2.0.CO;2, 1993.

Deser, C., Phillips, A. S., and Hurrell, J. W.: Pacific interdecadal climate variability: linkages between the tropics and the North Pacific during boreal winter since 1900, J. Clim., 17, 3109–3124, https://doi.org/10.1175/1520-0442(2004)017<3109:PICVLB>2.0.CO;2, 2004.

Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: the role of internal variability, Clim. Dyn., 38, 527–546, https://doi.org/10.1007/s00382-010-0977-x, 2012.

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. L.: Seasonal climate predictability and forecasting: status and prospects, WIREs Clim. Change, 4, 245–268, https://doi.org/10.1002/wcc.217, 2013.

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., and Robinson, N.: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world?, Geophys. Res. Lett., 41, 5620–5628, https://doi.org/10.1002/2014GL061146, 2014.

Gastineau, G. and Frankignoul, C.: Influence of North Atlantic SST variability on the atmospheric circulation, J. Clim., 28, 1396–1416, https://doi.org/10.1175/JCLI-D-14-00424.1, 2015.

Geng, X., Kug, J. S., and Kosaka, Y.: Future changes in wintertime ENSO–NAO teleconnection under greenhouse warming, npj Clim. Atmos. Sci., 7, 81, https://doi.org/10.1038/s41612-024-00627-z, 2024.

Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., Lamarque, J.-F., Fasullo, J. T., Bailey, D. A., Lawrence, D. M., and Mills, M. J.: High climate sensitivity in the Community Earth System Model version 2 (CESM2), Geophys. Res. Lett., 46, 8329–8337, https://doi.org/10.1029/2019GL083978, 2019.

Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in regional climate predictions, Bull. Amer. Meteor. Soc., 90, 1095–1107, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Sabater, J. M., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux,

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N.: The ERA5 global reanalysis. Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hoskins, B. J. and Karoly, D. J.: The steady linear response of a spherical atmosphere to thermal and orographic forcing, J. Atmos. Sci., 38, 1179–1196, https://doi.org/10.1175/1520-0469(1981)038<1179:TSLROA>2.0.CO;2, 1981.

Hurrell, J. W.: Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation, Science, 269, 676–679, https://doi.org/10.1126/science.269.5224.676, 1995.

Hurrell, J. W. and Deser, C.: North Atlantic climate variability: the role of the North Atlantic Oscillation, J. Mar. Syst., 79, 231–244, https://doi.org/10.1016/j.jmarsys.2009.11.002, 2010.

Hurrell, J. W., Kushnir, Y., Ottersen, G., and Visbeck, M.: An overview of the North Atlantic Oscillation, Geophys. Monogr., 134, 1–35, American Geophysical Union, Washington D.C., https://doi.org/10.1029/134GM01, 2003.

Hurrell, J. W. and van Loon, H.: Decadal variations in climate associated with the North Atlantic Oscillation, Clim. Change, 36, 301–326, https://doi.org/10.1023/A:1005314315270, 1997.

Jackson, L. C., Biastoch, A., Buckley, M. W., Desbruyères, D. G., Frajka-Williams, E., Moat, B., and Robson, J.: The evolution of the North Atlantic Meridional Overturning Circulation since 1980, Nat. Rev. Earth Environ., 3, 241–254, https://doi.org/10.1038/s43017-022-00263-2, 2022.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L.: Physics-informed machine learning, Nat. Rev. Phys., 3, 422–440, https://doi.org/10.1038/s42254-021-00314-5, 2021.

Kushnir, Y.: Interdecadal variations in North Atlantic sea surface temperature and associated atmospheric conditions, J. Clim., 7, 141–157, https://doi.org/10.1175/1520-0442(1994)007<0141:IVINAS>2.0.CO;2, 1994.

Kushnir, Y., Scaife, A. A., Arritt, R., Balsamo, G., Boer, G., Doblas-Reyes, F., Hawkins, E., Kimoto, M., Kolli, R. K., Kumar, A., Matei, D., Matthes, K., Müller, W. A., O'Kane, T., Perlwitz, J., Power, S., Raphael, M., Shimpo, A., Smith, D., Tuma, M., and Wu, B.: Towards operational predictions of the near-term climate, Nat. Clim. Change, 9, 94–101, https://doi.org/10.1038/s41558-018-0359-7, 2019.

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., Kröger, J., Takano, Y., Ghosh, R., Hedemann, C., Li, C., Li, H., Manzini, E., Notz, D., Putrasahan, D., Boysen, L., Claussen, M., Ilyina, T., Olonscheck, D., Raddatz, T., Stevens, B., and Marotzke, J.: The Max Planck Institute Grand Ensemble: enabling the exploration of climate system variability, J. Adv. Model. Earth Syst., 11, 2050–2069, https://doi.org/10.1029/2019MS001639, 2019.

Marshall, J., Kushnir, Y., Battisti, D., Chang, P., Czaja, A., Dickson, R., Hurrell, J., McCartney, M., Saravanan, R., and Visbeck, M.: North Atlantic climate variability: phenomena, impacts and mechanisms, Int. J. Climatol., 21, 1863–1898, https://doi.org/10.1002/joc.693, 2001.

Menary, M. B., Robson, J., Allan, R. P., Booth, B. B. B., Cassou, C., Gastineau, G., Gregory, J., Hodson, D., Jones, C., Mignot, J., Ringer, M., Sutton, R., Wilcox, L., and Zhang, R.: Aerosol-forced AMOC changes in CMIP6 historical simulations, Geophys. Res. Lett., 47, e2020GL088166, https://doi.org/10.1029/2020GL088166, 2020.

Peng, S., Robinson, W. A., and Li, S.: Mechanisms for the NAO responses to the North Atlantic SST tripole, J. Clim., 16, 1987–2004, https://doi.org/10.1175/1520-0442(2003)016<1987:MFTNRT>2.0.CO;2, 2003.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, J. Geophys. Res. Atmos., 108, 4407, https://doi.org/10.1029/2002JD002670, 2003.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, F.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Rodwell, M. J., Rowell, D. P., and Folland, C. K.: Oceanic forcing of the wintertime North Atlantic Oscillation and European climate, Nature, 398, 320–323, https://doi.org/10.1038/18648, 1999.

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., Eade, R., Fereday, D., Folland, C. K., Gordon, M., Hermanson, L., Knight, J. R., Lea, D. J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A. K., Smith, D., Vellinga, M., Wallace, E., Waters, J., and Williams, A.: Skillful long-range prediction of European and North American winters, Geophys. Res. Lett., 41, 2514–2519, https://doi.org/10.1002/2014GL059637, 2014.

Scaife, A. A. and Smith, D.: A signal-to-noise paradox in climate science, npj Clim. Atmos. Sci., 1, 28, https://doi.org/10.1038/s41612-018-0038-4, 2018.

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., Müller, W. A., Nicolí, D., Ortega, P., Pankatz, K., Pohlmann, H., Robson, J., Ruggieri, P., Sospedra-Alfonso, R., Swingedouw, D., Wang, Y., Wild, S., Yeager, S., Yang, X., and Zhang, L.: North Atlantic climate far more predictable than models imply, Nature, 583, 796–800, https://doi.org/10.1038/s41586-020-2525-0, 2020.

Sutton, R. T. and Hodson, D. L. R.: Atlantic Ocean forcing of North American and European summer climate, Science, 309, 115–118, https://doi.org/10.1126/science.1109496, 2005.

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., von Salzen, K., Yang, D., and Winter, B.: The Canadian Earth System Model version 5 (CanESM5.0.3), Geosci. Model Dev., 12, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019, 2019.

Visbeck, M., Hurrell, J. W., Polvani, L., and Cullen, H. M.: The North Atlantic Oscillation: past, present, and future, Proc. Natl. Acad. Sci., 98, 12876–12877, https://doi.org/10.1073/pnas.231391598, 2001.

Visbeck, M., Chassignet, E. P., Curry, R. G., Delworth, T. L., Dickson, R. R., and Krahmann, G.: The ocean's response to North Atlantic Oscillation variability, in: The North Atlantic Oscillation: Climatic Significance and Environmental Impact, edited by: Hurrell, J. W. et al., Geophys. Monogr., 134, 113–145, https://doi.org/10.1029/134GM06, 2003.

Walters, D., Baran, A. J., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., Furtado, K., Hill, P., Lock, A., Manners, J., et al.: The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 configurations, Geosci. Model Dev., 12, 1909–1963, https://doi.org/10.5194/gmd-12-1909-2019, 2019.

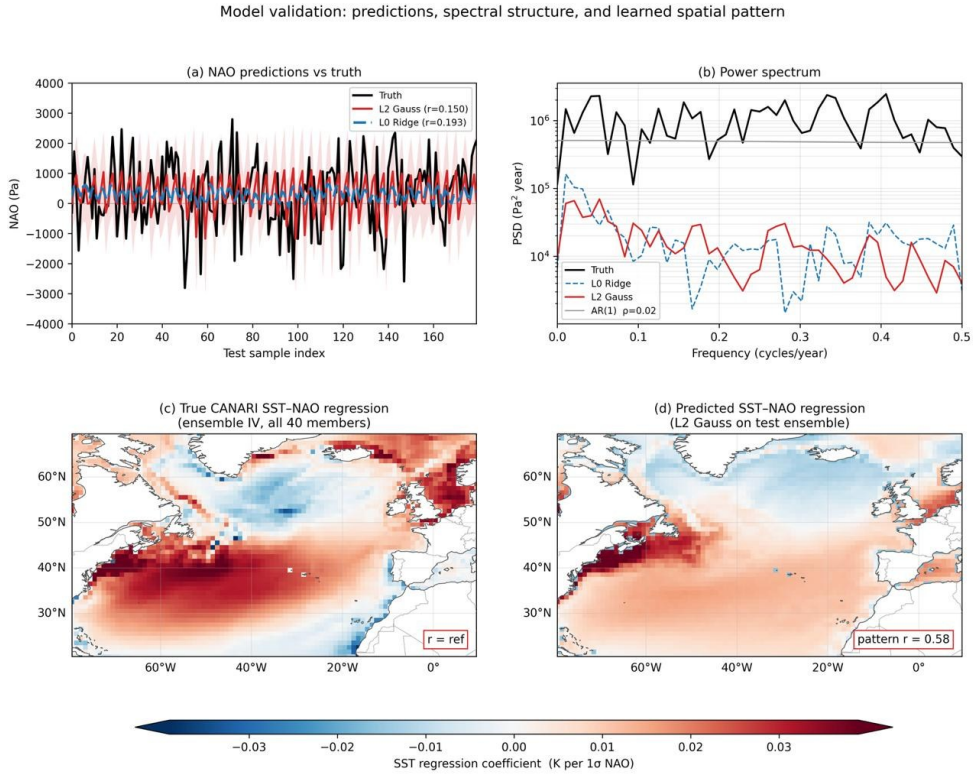
Weisheimer, A., et al.: The signal-to-noise paradox in climate forecasts: revisiting our understanding and identifying future priorities, Bull. Amer. Meteor. Soc., 105, E651–E659, https://doi.org/10.1175/BAMS-D-22-0067.1, 2024.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

670 Woollings, T., Hannachi, A., and Hoskins, B.: Variability of the North Atlantic eddy-driven jet stream, Q. J. R.
671 Meteorol. Soc., 136, 856–868, https://doi.org/10.1002/qj.625, 2010.
672

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

673 **Figures**



674

675 **Figure 1**. Model validation. (a) NAO predictions versus truth on the HIST held-out test set (2000–2014; 8
676 members). L0 Ridge (blue dashed, r = 0.193) and L2 Gaussian (red, r = 0.150). (b) Power spectral density of L0
677 Ridge predictions (blue), observed NAO (black), and AR(1) baseline (grey). (c) Reference CANARI SST–NAO
678 regression (ensemble internal variability, all 40 members; K per 1σ NAO). (d) Predicted SST–NAO regression
679 (L2 Gaussian, pattern r = 0.58). Stippling: $p < 0.05$.

680

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

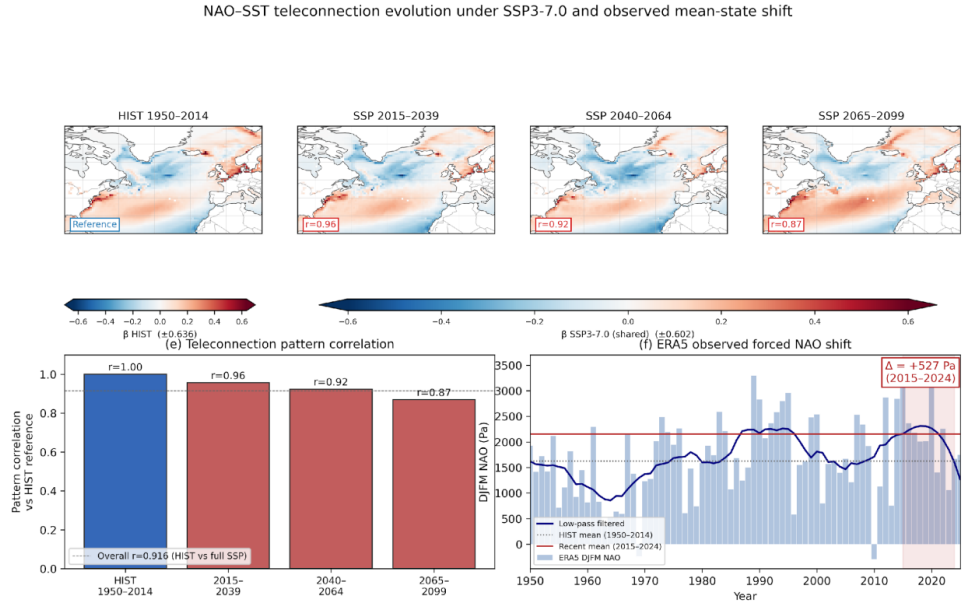


Figure 2. NAO–SST teleconnection evolution and observed shift. (a)–(d) DJFM NAO–SST regression maps for HIST 1950–2014 and successive SSP3-7.0 windows. Pattern correlations with the historical reference are shown as numbers on the bottom-left of each panel (r = 0.96, 0.92, 0.87). (e) Teleconnection pattern correlation vs HIST: overall r = 0.92 (HIST vs full SSP) - aggregate value across the full SSP period; distinct from the per-window pattern correlations 0.96, 0.92, 0.87 annotated in panels c–d. (f) ERA5 DJFM NAO index (1950–2024) (raw, un-normalised, values in Pa): HIST mean 1626 Pa (dotted), recent mean 2153 Pa (red line), forced shift Δ = +527 Pa.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

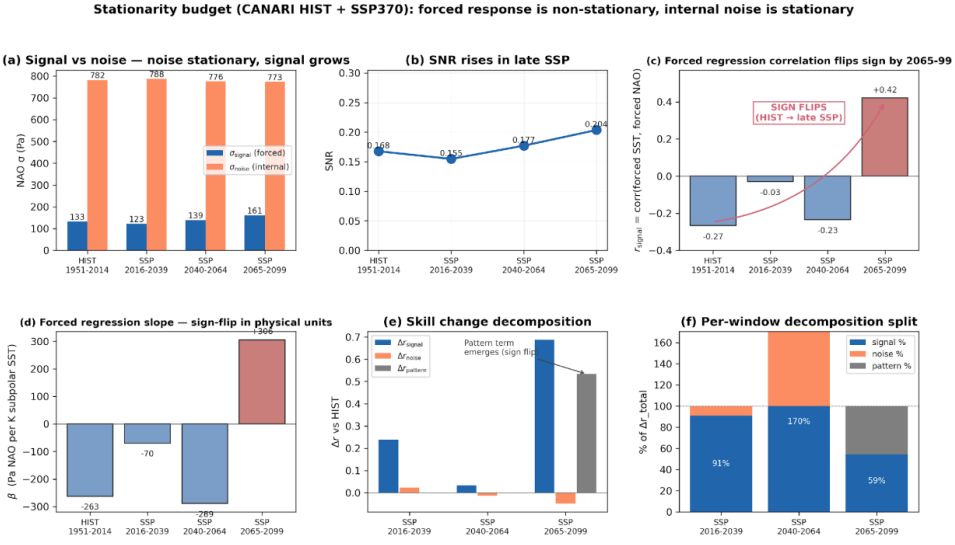


Figure 3. Three-term stationarity budget. (a) Forced NAO σ_signal (blue) and internal NAO σ_noise (orange; 782–773 Pa per period) and forced NAO σ_signal (133→123→139→161 Pa per period). (b) SNR (σ_signal/σ_noise). (c) Forced-regression correlation r_signal (d) Forced regression slope β. (e) Skill change decomposition (Δr): signal (blue), noise (orange), and pattern (grey) terms per SSP window. (f) Per-window fractional contributions of each term to Δρ_total (percentages shown as bar labels).

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

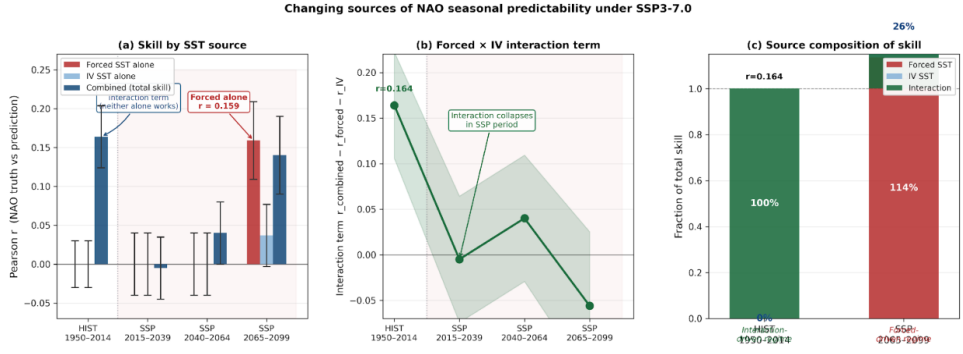


Figure 4. Changing sources of NAO seasonal predictability. (a) Pearson r achieved by the L0 hierarchy from forced SST alone (red), IV SST alone (light blue), and combined SST (dark blue) per window. Error bars: 95% bootstrap CI. (b) Interaction term r_interaction = r_combined − r_forced − r_IV per window. (c) Source composition of skill in HIST and 2065–2099. Total skill r annotated above bars.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

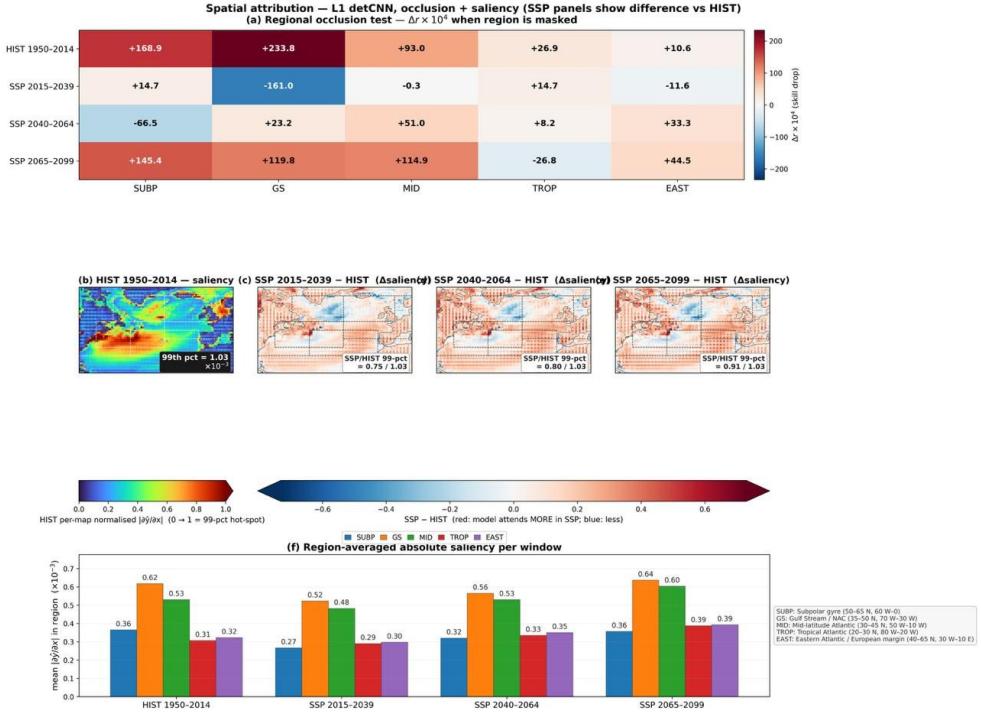


Figure 5. Spatial attribution of NAO seasonal predictability: regional occlusion and network saliency across the forcing period. (a) Regional occlusion test: skill loss Δr when each North Atlantic sub-region is masked from the L1 detCNN input, for four time windows (HIST 1950–2014; SSP 2015–2039; SSP 2040–2064; SSP 2065–2099). Positive Δr: region contributes to skill; negative Δr: region confounds prediction. Regions: SUBP = subpolar gyre (50–65°N, 60°W–0°); GS = Gulf Stream / North Atlantic Current (35–50°N, 75°W–30°W); MID = mid-latitude Atlantic (30–45°N, 50°W–10°W); TROP = tropical North Atlantic (20–30°N, 80°W–20°W); EAST = eastern Atlantic and European margin (40–65°N, 30°W–10°E). (b) HIST 1950–2014 per-map normalised absolute saliency $|\partial\hat{y}/\partial x|$, normalised so that the 99th-percentile hot-spot = 1 (value: 1.03×10^{-3}). (c–e) Saliency difference maps (SSP window minus HIST): red indicates regions to which the model attends more in the SSP period; blue indicates reduced attention. SSP/HIST 99th-percentile ratios are annotated. (f) Region-averaged absolute saliency per window.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

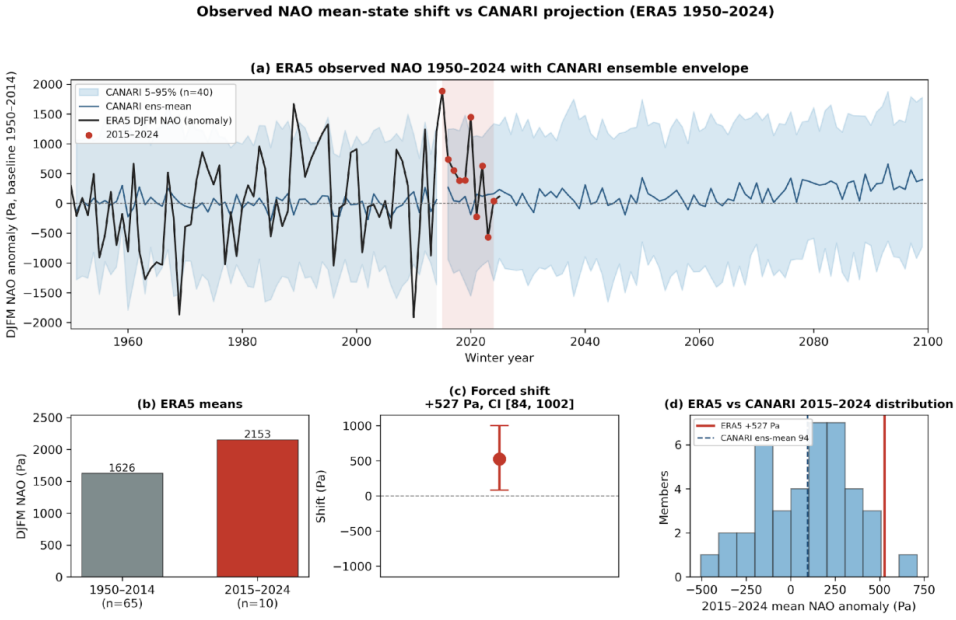


Figure 6. Observational evidence for forced nonstationarity. (a) ERA5 DJFM NAO anomaly (1950–2100, black) with CANARI 5–95% envelope (blue shading) and ensemble mean (blue line); 2015–2024 anomalies highlighted (red dots). (b) Period means: HIST 1950–2014 = 1626 Pa; recent 2015–2024 = 2153 Pa. (c) Bootstrap-estimated forced shift: +527 Pa (95% CI: 84–1002 Pa). (d) Distribution of CANARI 2015–2024 member mean NAO anomalies (N = 40) and ERA5 anomaly (red vertical line).

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

Appendix A: ML Hierarchy — Full Technical Details

This appendix specifies the prediction hierarchy in the order in which an input is processed. The training, validation and out-of-sample evaluation protocol are described in A1. A predictor SST field at each of K = 4 winter lags is first encoded by a shared-weight convolutional backbone (A2) into a feature vector per lag. The K = 4 lag features are then combined by a learned attention layer (A3) into a single time-attended feature vector. In parallel, a frozen physics layer (A4) projects the raw SST onto the canonical NAO–SST tripole and produces a single physics-informed scalar. The attended feature and the physics scalar are concatenated and passed to one of four output heads (A5) — ridge, deterministic, Gaussian, or Gaussian-mixture — which differ only at this last stage and which together constitute the four tiers L0, L1, L2, L3 used throughout the paper.

A1. Training, validation, and out-of-sample evaluation protocol. Models are fit on CANARI HIST members 1–32 over winters 1950–1999, with a within-HIST validation set of members 1–32 over 2000–2014 used only for early-stopping on the dual-loss objective (patience = 30 epochs). Held-out HIST test set: members 33–40 over 2000–2014 (1296 member × winter pairs). SSP3-7.0 evaluation: each model is applied without further training to all 40 members across each of the three SSP windows (2015–2039: 488 pairs; 2040–2064: 500 pairs; 2065–2099: 700 pairs). No fine-tuning, transfer learning, or recalibration is applied between HIST and SSP, and no SSP statistics enter training. Input standardisation uses the HIST training mean and standard deviation per grid point and per lag computed from members 1–32, 1950–1999, applied identically to the HIST test set and to all SSP windows. NAO targets are normalised by σ_{obs} = 774.9 Pa (ERA5 DJFM NAO standard deviation, 1950–2014) for both training and evaluation, so reported skill numbers are dimensionless and directly comparable across windows. The L0 ridge regularisation parameter is selected by 5-fold cross-validation within the HIST training fold; once selected, the L0 coefficients are frozen and applied to all evaluation windows, identically to L1–L3.

A2. Shared-weight Conv2D backbone. Each DJFM SST lag map passes through three convolutional layers (32→64→128 filters, 3×3, ReLU, BatchNorm, GlobalAveragePool). Weights are shared across K = 4 lag maps. Output: per-lag feature vectors $h_\ell \in \mathbb{R}^{128}$.

A3. Lag attention. Scores $e_\ell = w^T h_\ell + b$; weights $\alpha_\ell = \exp(e_\ell)/\Sigma \exp(e_\ell)$; attended feature $\hat{h} = \Sigma_\ell \alpha_\ell h_\ell$. Converged: α = [0.17, 0.09, 0.25, 0.49] (lags 1–4).

A4. Frozen physics layer. β_{phys} from ERA5–HadISST NAO–SST regression (1950–2014), L2-normalised, ∇ = 0 throughout. $z(t) = \beta_{phys} \cdot X(t) + b$. Concatenated: $[\hat{h}; z] \in \mathbb{R}^{129}$.

A5. Output heads and loss. L0: ridge on 10 SST PCs (extreme member target only; no dual loss, no frozen physics). L1: Dense(2) → (μ_t, μ_{t+1}). L2: Dense(4) → Gaussian (μ, σ). L3: Dense(8) → GMM K=2. Loss: $\mathscr{L} = (1-\lambda)\cdot\mathrm{MSE}(\hat{y}, y^*) + \lambda\cdot(\sigma(\hat{y}) - \sigma_{obs})^2$, λ=0.6, σ_{obs}=774.9 Pa (ERA5 DJFM NAO standard deviation, 1950–2014). Adam, lr=1×10^{-3} cosine decay, batch=32, 300 epochs, early stopping patience=30.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

Appendix B: Budget Derivation and Closure Verification

This appendix derives the three-term decomposition introduced in §2.3 and verifies that it closes additively across all SSP windows. The decomposition compares the forced-regression correlation between ensemble-mean SST and ensemble-mean NAO in HIST against the same quantity computed in each SSP window. Three independent contributions are identified: a signal term Δr_{signal} capturing change in the forced amplitude, a noise term Δr_{noise} capturing change in the SNR, and a pattern term $\Delta r_{pattern}$ capturing the reorganisation of the forced regression slope β. Their additive sum is required to match the directly computed Δr_{total} to within numerical precision; the residual is reported below.

$\Delta r_{total} = \Delta r_{signal} + \Delta r_{noise} + \Delta r_{pattern}$. $\Delta r_{signal} = r_{HIST} - r_{signal}(w)$, where $r_{signal}(w)$ = corr(ensemble-mean SST, ensemble-mean NAO) in window w, estimated after re-centring SSP SST inputs to the HIST climatological mean. Δr_{noise} captures the change in $\sigma_{signal}/\sigma_{noise}$ ratio. $\Delta r_{pattern}$ captures the sign reversal of the forced regression slope β. Closure verification: $|\Delta r_{signal} + \Delta r_{noise} + \Delta r_{pattern} - \Delta r_{total}| < 0.003$ in all SSP windows. Cross-correlation between ensemble-mean NAO and intra-ensemble NAO spread: $|\rho| \leq 0.03$, confirming approximate additivity of forced and internal contributions.

Architecture-invariance and noise-stationarity test statistics (referenced in §3.3). Pairwise Fisher z-test p-values exceed 0.49 across all (architecture, window) comparisons. The largest cross-architecture $|\Delta r|$ within any single SSP window is 0.037 (L0 vs L3) and 0.016 (L1 vs L3). The noise-term F-test for variance equality across windows gives $p = 0.40$, consistent with the σ_{noise} = 782–773 Pa stationarity reported in §3.3. The forced regression complements β with the correlation coefficient $r_{signal} = -0.27$ in HIST and +0.42 in 2065–2099, and σ_{signal} evolves through 133, 123, 139 and 161 Pa across HIST, 2015–2039, 2040–2064 and 2065–2099 respectively (Fig. 3a).

Appendix C: Spatial Attribution — Regional Occlusion and Gradient Saliency

C1. Overview and motivation

Two complementary attribution techniques diagnose which regions of the North Atlantic SST field carry predictive NAO information, and how this spatial footprint evolves across historical and SSP3-7.0 windows. Regional occlusion evaluates prediction skill when each sub-region is systematically masked — a model-agnostic measure of regional importance. Gradient saliency computes the sensitivity of the L1 detCNN output to each input grid point — a spatially continuous attention map. Together they address whether the spatial origin of NAO predictability shifts as the forced-predictability regime establishes itself.

C2. Regional definitions

Five mutually exclusive sub-regions partition the North Atlantic predictor domain (0–80°N, 90°W–20°E): SUBP — subpolar gyre (50–65°N, 60°W–0°); GS — Gulf Stream / North Atlantic Current (35–50°N, 75°W–30°W); MID — mid-latitude Atlantic (30–45°N, 50°W–10°W); TROP — tropical North Atlantic (20–30°N, 80°W–20°W); EAST — eastern Atlantic and European margin (40–65°N, 30°W–10°E). The EAST region encompasses the seas immediately west of the United Kingdom and the European continental shelf — the SST sector most directly coupled to NAO-driven temperature and precipitation anomalies over the British Isles. Regions are held fixed across all time windows.

https://doi.org/10.5194/egusphere-2026-2686
Preprint. Discussion started: 26 May 2026

C3. Regional occlusion test

For each region R and time window w, SST values within R are replaced by their HIST climatological mean (1950–2014 ensemble mean) before passing the input through the trained L1 detCNN. The occlusion skill loss is:

$$\Delta r\_occlusion(R, w) = r\_full(w) - r\_occluded(R, w) \quad (C1)$$

A positive value indicates region R contributes positively to skill; negative indicates R confounds prediction. The sign reversal at GS in 2015–2039 ($\Delta r = -0.016$) indicates that Gulf Stream SST variance projects more onto noise than signal during the forced regression transition ($\beta \approx -70$ Pa K^{-1}). Uncertainty is estimated by block bootstrap (1,000 resamplings; 95% CI ≈ ±0.008–0.013 across windows). Values $|\Delta r\_occlusion| < 0.005$ are not distinguished from zero at the 95% level.

C4. Gradient saliency analysis

The mean absolute saliency over all test-set samples and K = 4 lags at grid point i is:

$$S\_bar(i,w) = (1/N\_w\ K)\ Sum\_\{t\ in\ w\}\ Sum\_\{l=1\}\^\{K\}\ |dy\_hat(t)/dx\_i(t,l)| \quad (C2)$$

Gradients are computed via backpropagation (PyTorch autograd) through frozen network weights. The HIST saliency map (Fig. 5b) is normalised by its 99th-percentile value q_{0.99} = 1.03 x 10^{-3} so the hottest 1% of HIST grid points have normalised saliency ~1. The SSP/HIST 99th-percentile ratios rho(w) = q_{0.99}(S_bar_SSP(w)) / q_{0.99}(S_bar_HIST) reported in Fig. 5c-e reflect changes in peak attention magnitude: rho = 0.75 in 2015-2039, recovering to 0.91 by 2065-2099. Saliency difference maps (Fig. 5c-e) show Delta_S_bar(i,w) = S_bar(i,w) - S_bar(i,HIST): red = increased attention in SSP; blue = decreased. Region-averaged absolute saliency (Fig. 5f) is the spatial mean over each region, excluding land and sea-ice points. Saliency captures first-order local sensitivity; occlusion captures the integrated skill change including higher-order cross-region effects. Both use the L1 detCNN to allow spatial maps; the source decomposition in Section 2.4 uses L0 for the reasons given there.

C5. Reproducibility

All computations use scripts spatial_attribution.py and occlusion_test.py at [GitHub DOI], random seed 42. Preprocessing applies HIST training-period statistics identically to HIST and SSP inputs. Regional mask arrays are at data/regional_masks.nc in the repository.