Review of: *"Forced reorganisation of the North Atlantic SST–NAO relationship and implications for European winter predictability"* by Ioana Colfescu, submitted to WCD, Ref. MS No.: egusphere-2026-2686

General assessment

This manuscript investigates how the relationship between North Atlantic SST anomalies and the winter NAO may evolve under anthropogenic forcing. Using the 40-member CANARI ensemble under historical and SSP3-7.0 forcing, the authors separate ensemble-mean and ensemble-departure components of SST and NAO, and combine this decomposition with a hierarchy of statistical and machine-learning prediction models. They conclude that the internally generated teleconnection remains broadly stable through the twenty-first century, while the externally forced component weakens and eventually reverses sign, producing a transition in the sources and magnitude of seasonal NAO predictability.

The question is interesting and potentially important. Diagnosing how the SST–NAO relationship and the mean state on which it operates evolve under external forcing is a valuable exercise, and a large ensemble is well suited to the task, and I appreciate the use of several architectures as a robustness test.

My main concern is that the manuscript moves freely between distinct concepts: forecast skill, predictability, SST–NAO correlation, externally forced covariance, SST-forced atmospheric response, which are not equivalent but are at times treated as such. This makes it hard to determine what is being decomposed and what the diagnostics imply for real prediction systems, and the same difficulty affects the stationarity budget, the training target and the predictability-source decomposition. Furthermore, the central result (the sign reversal of the forced SST–NAO regression) is presented without any uncertainty estimate, and is computed from ensemble-mean series whose variance may be dominated by residual sampling noise.

For these reasons, and on the basis of the remarks that follow, I recommend **major revision**.

Major comments

1. Forecast skill, model-world predictive skill and predictability need to be clearly distinguished

CANARI is not an initialized seasonal prediction system, and the "predictions" analysed here come from the statistical/ML models L0–L3, using simulated SSTs to predict simulated NAO. This is a legitimate framework for studying model-world SST-based potential predictability, but it is not equivalent to evaluating an operational forecast system. In the SSP periods the verification target is necessarily the simulated CANARI NAO, yet the manuscript refers to "predicted and observed NAO"; this should be revised.

I recommend distinguishing explicitly, throughout, between r_pred = corr(NAO_predicted, NAO_CANARI), which may reasonably be described as out-of-sample model-world prediction skill; r_SST-NAO = corr(SST, NAO), which characterises an SST–NAO relationship or potential predictability source; and the spatial pattern correlations used to compare regression maps. At present all three are denoted by r, which I found confusing.

More generally, I suggest reserving "forecast skill" for the performance of a specified model against an independent target, and "predictability" for properties of the underlying signal. A correlation between ensemble-mean SST and ensemble-mean NAO should not be called forecast skill without further justification.

This matters because much of the discussion extends the results directly to operational systems and to future European winter prediction. The paper would be stronger if framed as diagnosing North-Atlantic-SST-related potential predictability in the CANARI model world, with the implications for initialized systems discussed separately.

2. The three-term stationarity budget is difficult to follow.

I found Section 2.3 the most difficult part of the manuscript to follow. The budget is said to separate "forecast-skill changes", but the r it uses appears to be a correlation between ensemble-mean SST and ensemble-mean NAO, not between predicted and target NAO. It is therefore unclear in what sense Δr_total, Δr_signal, Δr_noise and Δr_pattern decompose forecast skill. The ambiguity is explicit in one place: line 341–342 states that L0 "is used as the baseline for the stationarity budget in Sect. 3.3", whereas Appendix B

defines every budget quantity as a correlation between ensemble-mean SST and ensemble-mean NAO (lines 762–764, 768–770), with no reference to any (L) model output. If the budget is indeed computed from ensemble-mean series alone, line 341–342 is misleading and should be corrected; if L0 does enter the calculation, Appendix B needs to say how

Δr_noise raises a further difficulty. Line 291–293 defines r_noise(w) as r(w) recomputed after rescaling the SSP NAO variance to the HIST values of σ_signal and σ_noise. The nature of this rescaling is not specified, and its effect depends on what is left unstated. Please give the explicit operation applied to the NAO series and state which quantities the correlation is computed from in this term.

More generally, I recommend providing an explicit mathematical definition of every r used, the complete equations for all budget terms with a statement of which quantities each is computed from, a precise description of the counterfactual SST and NAO fields used in each case, and ideally a schematic. "Forced regression amplitude", "internal variability amplitude" and "spatial reorganisation of the forced regression" should be defined before the budget is introduced.

Please also clarify how Δr_pattern is obtained. Line 294–295 describes it as what remains once the amplitude and noise contributions have been removed, which is the definition of a residual, while lines 296–297 state that the three terms are computed independently from their own estimators and are not derived sequentially. These cannot both hold, and no formula for Δr_pattern is given in either the main text or Appendix B. If the term is a residual, the reported closure is true by construction and should not be presented as validation.

Finally, the paper contains two distinct three-term decompositions, both introduced in Sect. 2.3: the stationarity budget and the predictability-source decomposition. They are computed from different quantities and answer different questions, but share a three-term structure, the symbol r, and terms that can be read as "forced" and "internal" in both, with different meanings in each. Distinct names, and a sentence stating how the two differ, would help the reader considerably.

3. No uncertainty is reported for the central result

The sign reversal of the forced regression is probably the paper's main claim, but I could find no confidence interval for r_signal, β or any budget term, in any window. Sect. 2.3 describes bootstrap and block resampling, yet no resulting uncertainty appears anywhere. This matters because the budget quantities are computed on ensemble-mean series of 25 to 65 winters that are strongly trended and autocorrelated, so the effective degrees of freedom are few.

A related gap concerns the models. Sect. 3.7 tests whether the configurations differ from one another, but the claim in the text is that skill declines across windows, and I could not find that comparison tested. The available numbers (HIST r = 0.19, 95% CI 0.11–0.27; late century r ≈ 0.16, with line 416 noting that apparent skill is similar in the two periods) suggest the decline may not be distinguishable.

Please report confidence intervals for r_signal, β and the budget terms in every window, using an effective sample size that accounts for autocorrelation, and state whether the sign reversal is distinguishable from no change. A table of skill by configuration and window, with confidence intervals, would address the second point.

4. Residual sampling noise in the ensemble mean may be comparable to the forced signal

Every budget quantity is computed from ensemble-mean series, so it matters how completely 40 members remove internal variability. Averaging a finite ensemble leaves a residual of order σ_noise/sqrt(N), and given the reported σ_noise this residual is of the same order as the reported σ_signal. A substantial part of the ensemble-mean NAO variance may therefore be residual internal variability rather than forced response, with the true forced amplitude and SNR both smaller than quoted.

This offers an alternative reading of the central result: if the ensemble-mean series carry a large noise component, a correlation between them can change from window to window for reasons unrelated to forcing. The manuscript notes that 30–40 members give robust forced estimates (lines 175–176) but does not address how the residual compares with the forced amplitude here. I recommend reporting a noise-corrected estimate of σ_signal, and recomputing β, r_signal and the budget terms on reduced-size sub-

ensembles to compare the spread across sub-ensembles with the window-to-window changes attributed to forcing.

5. "Forced" is ambiguous, and the causal interpretation of the ensemble-mean relationship needs reconsidering

"Forced component" is used in two senses: the externally forced component of SST or NAO, estimated from the ensemble mean; and the atmospheric response forced by SST anomalies. These differ conceptually: internally generated SST anomalies can force an atmospheric response, while an externally forced response can affect NAO variability without being mediated by North Atlantic SSTs. Please distinguish, for example, "externally forced SST/NAO component" from "SST-forced atmospheric response".

A relationship between ensemble-mean SST and ensemble-mean NAO establishes that the forced components covary. It does not establish the pathway external forcing → North Atlantic SST → NAO, since both may respond partly independently to the same forcing. The lagged mechanism documented in the literature (Rodwell et al., 1999; Czaja and Frankignoul, 2002; Gastineau and Frankignoul, 2015) concerns internally generated SST anomalies preceding the winter circulation and provides a sound basis for the internal component of this analysis, but it does not transfer automatically to the forced component. If the authors intend to argue for oceanic mediation, further evidence is needed.

Two missing specifications make it harder to assess whether the forced correlation reflects oceanic mediation. First, β requires ⟨SST⟩ to be a scalar, and therefore an area average, but the domain over which it is computed is never stated. The only indications are the axis labels of Fig. 3d ("per K subpolar SST") and Fig. S1c ("per K SST"), which disagree with each other. The only spatial domain defined explicitly in the manuscript (20–70°N, 80°W–10°E, l. 160–162) is the predictor field for the L0–L3 models, which is a different object. Since the paper describes the forced warming as spatially differential (subpolar and tropical nodes evolving in opposite directions) the value of β, and possibly its sign, may depend appreciably on this choice. Second, it is unclear whether the forced regression is lagged or contemporaneous. Please state the domain, season and lag of <SST> before introducing β.

I also suggest avoiding the implicit identification of the ensemble mean with the anthropogenic signal: it estimates the response to all external forcings, and over the historical period includes volcanic and solar forcing as well as residual internal (oceanic) variability.

6. The extreme-member training target needs justification

At each timestep the authors select the member whose NAO departs most strongly from the ensemble mean, arguing that this preserves the direction of the forced signal (lines 248–251). The text does not state whether the departure is taken in absolute or in signed terms, and the two cases have different implications.

If selection is based on absolute departure, the largest departure can be of either sign, and since internal departures are much larger than the forced signal, the sign of the selected target would often be opposite to that of the ensemble mean. In that case the direction of the forced signal is not preserved but obscured. If selection is signed, the stated claim follows, but the procedure then conditions the training target on the quantity being predicted. In either case, in fact, systematically selecting the most extreme outcome changes the distribution of the training target, which is no longer sampled from the ordinary distribution of member NAO values, and could introduce selection bias.

Please state the selection rule explicitly, and which SST field is paired with the selected target. This is relevant because if the SST of the same member is used, the training sequence jumps between members at each timestep, which breaks the multi-month ocean memory the predictors are meant to exploit; if a different SST field is used, input and target are no longer physically consistent. Please also show the resulting target distribution against the ordinary member distribution, and test sensitivity to alternatives such as variance rescaling of the ensemble-mean signal or random member-wise selection. The main conclusions should ideally be shown robust to this choice.

7. The predictability-source decomposition in Fig. 4 is not an additive attribution of skill

The interaction term is defined approximately as r_interaction = r_combined − r_forced − r_IV and interpreted as skill arising from the joint structure of the two SST components. Pearson correlations are not

additive: even with a linear L0 and ŷ_combined = ŷ_forced + ŷ_IV, it does not follow that corr(y, ŷ_combined) = corr(y, ŷ_forced) + corr(y, ŷ_IV). The quantity is a residual diagnostic, not the fraction of skill physically attributable to an interaction. Since strong statements follow (historical skill as interaction-driven, late-century skill as predominantly forced) I suggest either reformulating with a genuinely additive covariance or variance decomposition, or weakening the interpretation.

A related concern applies to the values themselves. The forced SST field is the same for every member in a given winter, so a prediction based on forced SST alone takes a single value per winter, while the target varies strongly between members. Correlating the two across pooled member–winter pairs must then give a low value, because most of the variance in the target has no counterpart in the prediction. A value of r_forced close to 0 in the historical period may therefore follow partly from the evaluation design rather than from the physics, which would weaken the inference that historical predictability is interaction-driven.

Finally, L0 is chosen here for its linearity, yet its learned spatial predictor bears little resemblance to the canonical tripole, which the more constrained configuration reproduces much better. It is not obvious that L0 is the right model from which to infer a physical decomposition of predictability sources; please justify the choice and repeat the decomposition with L2 as a check.

8. The operational implications, especially the recalibration claim, need qualification

The manuscript argues that a changing SST background will require systems calibrated on the historical climate to be recalibrated. This should distinguish statistical from initialized dynamical systems. A statistical SST–NAO model trained on a historical relationship must indeed be updated if that relationship changes. A dynamical system does not explicitly learn the historical regression; in principle the response to a changed SST background emerges from the coupled dynamics. It's true that dynamical systems require bias and drift correction against a fixed hindcast climatology, and those may need to evolve as the model climatology changes. But that is a different matter from retraining a statistical mapping.

I suggest distinguishing between statistical retraining of an empirical SST–NAO relationship; updating a climatological reference and bias correction for dynamical forecasts; and structural modification of the forecast model itself. The current wording blurs the three.

The experiments also deliberately retain HIST training and normalisation in future periods. If periodic recalibration is the proposed operational solution, it can be tested directly: repeat the future experiments with an updated climatological reference or a retrained model and quantify how much skill is recovered. Without that experiment the recommendation is more speculative than the wording suggests.

9. Historical non-stationarity should be discussed and tested more explicitly

The analysis is motivated partly by relatively high NAO skill reported in earlier studies, but such values often apply to specific periods while others show much lower skill; Weisheimer et al. (2020) emphasise strong multidecadal variations in NAO skill and predictability. This is directly relevant to a paper about stationarity. The introduction should make clear that historical NAO predictability is itself non-stationary and that the quoted values do not characterise the whole record.

I would also encourage a direct test within CANARI: compute the diagnostics over moving 20–30-year historical windows, or at least over historical windows of the same length as the SSP periods, to establish whether the projected changes lie outside the range of internally generated historical modulation. This is also relevant because the current comparison uses periods of different lengths; the correlation is not biased simply by window length, but its sampling uncertainty and sensitivity to low-frequency variability differ.

The proposed mechanism does not necessarily require a monotonic reduction in skill throughout 1950–2014 (other sources of predictability could produce strong historical modulation), but precisely for that reason the future transition should be evaluated relative to the amplitude of historical non-stationarity.

10. The physical interpretation of the sign reversal is stronger than the diagnostics shown

The manuscript states that broadly uniform North Atlantic warming weakens the subtropical-to-subpolar pressure gradient and produces a negative forced NAO tendency. A spatially uniform warming would not by itself weaken a meridional SST gradient, so the chain from warming to pressure-gradient change needs to be set out explicitly.

Similarly, AMOC weakening is said to drive the late-century differential warming and reverse the forced NAO response. This is plausible and consistent with previous literature, but I do not see a direct AMOC diagnostic establishing ΔAMOC → ΔSST → Δpressure gradient → ΔNAO. If the mechanism is not diagnosed in CANARI, I suggest presenting it as an interpretation consistent with previous studies rather than as demonstrated here; showing the CANARI AMOC index and the tropical-minus-subpolar SST gradient across the four windows would strengthen the case considerably. I would also note an apparent inconsistency with Sect. 3.5, where the tropical North Atlantic contributes negligibly to skill in all periods, although tropical warming drives the reversal in the narrative.

11. Recent ERA5 behaviour is consistent with the projected transition, but is not evidence of forced non-stationarity

Sect. 3.6 finds a positive shift of about +527 Pa in the 2015–2024 ERA5 NAO mean and interprets it as observational evidence for the forced transition. I think this is too strong. A ten-winter period from a single realisation can differ from the ensemble because of a stronger real forced response, residual internal variability, model biases in the forced response, differences in simulated internal variability, or some combination. The analysis shows that recent behaviour is consistent in sign with the projection, but does not attribute it to forced non-stationarity.

I recommend replacing "observational evidence for forced nonstationarity" with something more cautious, such as "recent observations are consistent with the sign of the projected forced transition"; the manuscript already acknowledges later that ten years is close to the lower limit for attribution.

12. The broader context of NAO predictability should be acknowledged

The focus on North Atlantic SSTs is reasonable, but several statements refer generally to "NAO predictability" or "European winter seasonal predictability". North Atlantic SSTs are only one source: recent work identifies tropical Indian and Pacific heating as drivers of NAO seasonal predictability (Senan et al., 2024, JGR-Atmosphere), and documents pronounced intermittency in NAO forecast skill, with well-forecast years associated with strong tropical forcing and poorly forecast years with weak ENSO forcing (Baker et al., 2024, Geophys. Res. Lett.). This is also relevant to the historical non-stationarity discussed above.

Please clarify throughout that the study diagnoses the North-Atlantic-SST-related component of NAO predictability rather than the total, and briefly discuss other sources.

Specific and minor comments

Line 26. Persistence alone does not provide predictability. It arises from persistence together with a robust relationship between antecedent SST and subsequent NAO.

Lines 69–70. The quoted r ≈ 0.6 should be qualified as applying to particular hindcast periods. These values also come from initialized dynamical systems verified against observations, and so are not directly comparable with the r ≈ 0.19 reported here for pooled member–winter correlations against a simulated target.

Line 82. The climatological SST background can also change through low-frequency internal variability, not only anthropogenic forcing.

Lines 110, 120–122. "Forced" is used in two senses; statements such as "a contraction of the forced regression toward zero" (line 111) are ambiguous between them. See also Major comment 5.

Lines 126–135. Please define "stationarity budget", "forced regression amplitude", "internal variability amplitude", "spatial reorganisation" and "complementary predictability-source analysis" before using them. Equations would help. See also Major comment 2.

Lines 154, 170. The ensemble mean captures all external forcings, not the anthropogenic signal alone. On the causal interpretation, see also Major comment 5.

Lines 186–188. If the externally forced component accounts for only about 3% of total NAO variance, it would help to state early on why changes in this component matter for seasonal prediction.

Line 193. "Forecast skill" is used here but defined only at l. 241–242.

Lines 193–194. The evaluation periods have different lengths; please discuss the consequences for sampling uncertainty and provide an equal-length-window sensitivity test.

Line 198. Please clarify "ensemble mean of member-wise regressions": is the regression computed separately for each member and the resulting maps then averaged? The relationship between this estimator and the statement at lines 278–281, that ensemble-averaged member-wise regressions are not used for the budget because they carry an internal-variability bias, is also unclear. The two passages appear to describe the same estimator applied to different objects, but as written it is easy to read them as contradicting each other. A sentence explaining which estimator is used where, and why, would remove the ambiguity.

Line 216. A one-line definition of ridge regression would help, since L0 underpins both Sect. 3.3 and Sect. 3.4.

Lines 223–256. Lines 223–256. The description of the physical constraint is inconsistent: line 229 attributes it to L3, line 253 to L2 and L3. A compact table of L0–L3 would help: input, architecture, output distribution, loss, and whether the frozen constraint is used. Appendix A would also benefit from a table of notation: several symbols are introduced only through the formulae, and in A5 in particular $\hat{y}$ and y^* are not defined.

Line 256: "at low signal-to-noise" should read "at low signal-to-noise ratios."

Line 238. Please define "normalised using HIST training statistics" in the main Methods rather than only in the appendix.

Lines 241–242. Replace "observed NAO" with "target CANARI NAO" or "simulated NAO".

Line 267. "Reduced predictability" anticipates a result that is only demonstrated later. "Changes in predictability" would be more appropriate here.

Lines 323–332. Please show where the skill of every L0–L3 configuration is reported; Fig. 1 shows only a subset. "Truth" should be defined explicitly as the CANARI NAO.

Lines 336–343. L0 achieves its skill through variance modes that do not resemble the tripole (pattern correlation =−0.05), yet underpins both the budget baseline and the source decomposition. Please justify this choice, and consider repeating the decomposition with L2 as a robustness check, given that it is the configuration reported to recover the physical tripole.

Lines 344–348. The predicted spectra lie one to two orders of magnitude below the target in Fig. 1b, so "closely follows" can only refer to shape. The inference that the model exploits SST variability rather than NAO persistence also does not follow, since the models receive only SST as input when making a prediction and have no access to the NAO.

Lines 355–357. I would avoid describing the ocean-heat-content → atmosphere pathway as intrinsically an "internal-variability" pathway. SST/ocean anomalies can influence the atmosphere whether their origin is internal or externally forced. The distinction should refer to the origin of the SST anomaly, not to whether SST can dynamically force the atmosphere.

Lines 359–362. A pattern correlation is a domain-wide scalar and cannot localise where attenuation occurs. Please support the claim of subpolar concentration with a node-by-node quantification.

Lines 369–375. This ERA5 paragraph interrupts the progression between the teleconnection analysis and the budget, and duplicates Sect. 3.6. Consider moving it.

Line 378. "Sect. 3.6" appears to be a typo for Sect. 3.7. More importantly, the decline across architectures is referenced to Fig. 3, which shows the budget rather than per-configuration skill; that comparison does not appear anywhere.

Figure 3. Please check the SSP window labels: the figure and caption use 1951–2014 and 2016–2039 where the text uses 1950–2014 and 2015–2039; the same applies to Fig. S1. The six panels are also dense and hard to read at printed size.

Lines 381–382. Fig. S2 shows internal subpolar SST variance, not σ_noise, for which it is cited.

Line 392: The statement that the forced component of SST "has temporarily lost its coherent sign" is hard to interpret. If the meaning is that the forced regression coefficient weakens toward zero and then changes sign, please say so directly.

Lines 402–419 and Fig. 4. The values at l. 411–413 cannot be read from the panels; please annotate them or give a table. Please clarify what "skill by SST source" and panel (c) show, and why the stacked fractions exceed

100%; given Major Comment 7, "composition" may be misleading. Note also that r_combined in HIST is 0.16 here but L0 HIST skill is 0.19 in Sect. 3.1.

Lines 430–437. See Major Comment 9.

Appendix A. Sample sizes do not reconcile with the stated periods and member counts: members 33–40 over 2000–2014 give 120 member–winter pairs, not 1296; 40 members over 2015–2039 give 1000, not 488.