AGU100 ADVANCING EARTH AND SPACE SCIENCE

Geophysical Research Letters

RESEARCH LETTER

10.1029/2019GL083855

Key Points:

- Signal-to-noise paradox is a signature of an underconfident, unreliable prediction system
- CMIP5 models are unable to reproduce observed persistence of surface atmospheric temperature
- In CMIP5 models 70% of the globe exhibits the signal-to-noise paradox in SAT

Supporting Information:

- Supporting Information S1

Correspondence to:
F. Sévellec,
florian.sevellec@univ-brest.fr

Citation:
Sévellec, F., & Drijfhout, S. S. (2019). The signal-to-noise paradox for interannual surface atmospheric temperature predictions. *Geophysical Research Letters*, *46*, 9031–9041. https://doi.org/10.1029/2019GL083855

Received 12 OCT 2018
Accepted 5 JUL 2019
Accepted article online 22 JUL 2019
Published online 7 AUG 2019

The Signal-to-Noise Paradox for Interannual Surface Atmospheric Temperature Predictions

F. Sévellec[1,2] and S. S. Drijfhout[2,3,4]

[1]Laboratoire d'Océanographie Physique et Spatiale, Univ.-Brest CNRS IRD Ifremer, Brest, France, [2]Ocean and Earth Science, University of Southampton, Southampton, UK, [3]Koninklijk Nederlands Meteorologisch Instituut, De Bilt, Netherlands, [4]Institute for Marine and Atmospheric Science, University of Utrecht, Utrecht, Netherlands

Abstract The “signal-to-noise paradox” implies that climate models are better at predicting observations than themselves. Here, it is shown that this apparent paradox is expected when the relative level of predicted signal is weaker in models than in observations. In the presence of model error, the paradox only occurs in the range of small signal-to-noise ratio of the model, occurring for even smaller model signal-to-noise ratio with increasing model error. This paradox is always a signature of the prediction unreliability. Applying this concept to noninitialized simulations of Surface Atmospheric Temperature (SAT) of the CMIP5 database, under the assumption that prediction skill is associated with persistence, shows that global mean SAT is marginally less persistent in models than in observations. However, at a local scale, the analysis suggests that ~70% of the globe exhibits the signal-to-noise paradox for local SAT interannual forecasts and that the Signal-to-Noise Paradox occurs especially over the oceans.

1. Introduction

There is an increasing demand for accurate and reliable climate predictions on time scales from seasons to decades. This has led to the development of operational climate prediction systems with multiple models (Meehl et al., 2015), which allow for skillful seasonal predictions of hydrology (Svensson et al., 2015), energy supply (Clark et al., 2017), transport system disruption (Palin et al., 2016), and hurricane activity (Smith et al., 2010).

However with these promising results came along a somehow unexpected property: Climate prediction systems can be more accurate at predicting the real climate than predicting themselves (Kumar et al., 2014). Scaife et al. (2014) found that seasonal North Atlantic Oscillation predictions are inclined to develop this apparent paradox. Eade et al. (2014) also described this paradox for interannual predictions of Surface Atmospheric Temperature (SAT), mean sea level pressure, and precipitation in decadal predictions from DePreSys (Decadal Prediction System from the UK MetOffice, Smith et al., 2007). Similarly, Dunstone et al. (2016) verified its existence in interannual predictions of the North Atlantic Oscillation index. This behavior has been interpreted as a higher level of unpredictable components being present in the model than in observations (Siegert et al., 2016). Using an independent operational prediction system (PROCAST), Sévellec and Drijfhout (2018) observed the same property in interannual predictions of global-mean SAT. This discrepancy between model and real world prediction capability has been named the signal-to-noise paradox (see review by Scaife & Smith, 2018, for further details). To explain the paradox, a range of hypotheses has been put forward, such as the nonstationarity and the sampling uncertainty of predictions (Weisheimer et al., 2019).

2. Idealized Statistical Model

2.1. Definition and Prediction Accuracy Metrics

In this study we rationalize this behavior using an idealized statistical model. For that we consider a stochastic model for pseudo-observations (o) and model variables (m_i, where i denotes the index of ensemble members). Both have strictly identical statistics (following a centered and normal distribution with a standard deviation of 1), so no difference can be drawn from their statistical behavior (Figure 1), which avoids trivial conclusions regarding the signal-to-noise paradox (Boer et al., 2019). The pseudo-observations are split in two components: a predictable part (p) and a noise part (n_o). The model is also split in two components: A predictable part and a noise part (n_i, which differs for all model ensemble members). The

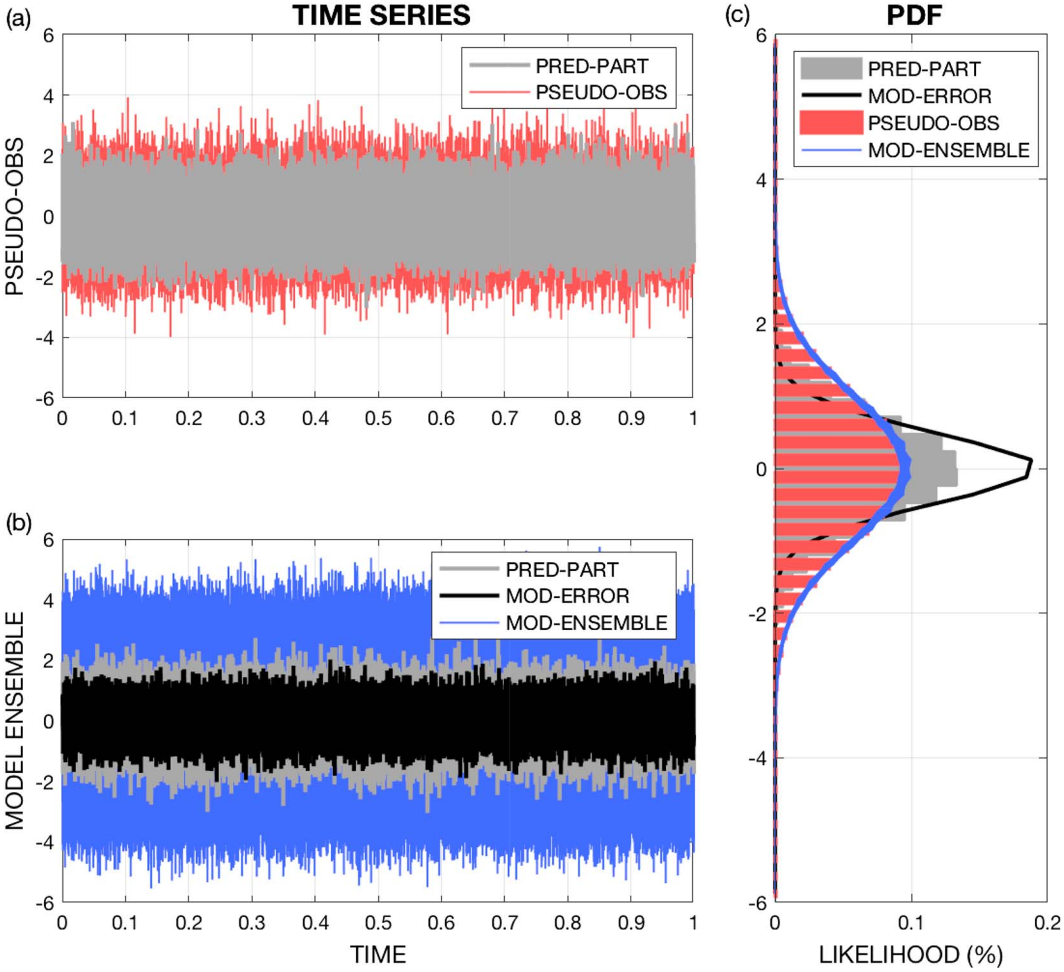

Figure 1. Model statistics. (a) Time series of pseudo-observations composed from a predictable component (gray) and noise. (b) Time series of the model ensemble composed from a predictable component (gray) and noise, the former having a systematic error (black). (c) Probability density function for (black) systematic error and (gray) predictable components, as well as (red) pseudo-observations and (blue) the model ensemble members. By construction all ensemble members and the pseudo-observations have strictly identical statistics (regardless of the level of systematic model error or predictable components). To get accurate statistics 50,000 time iterations and 4,000 members for the ensemble were used.

predictable part of the model is composed of two parts: a part common with the pseudo-observations (p) and a systematic model error (e, which is common to all model ensemble members). All parts (p, e, n_o, and n_i) consist of centered, normally distributed random variables. The pseudo-observations and model time series read as follows:

$$o(t) = A_o\left(\alpha_o p(t) + n_o(t)\right), \tag{1}$$

$$m_i(t) = A_m\left[\alpha_m A_p\left(\beta e(t) + p(t)\right) + n_i(t)\right], \tag{2}$$

respectively, where t is time; A_o, A_m, and A_p are the amplitudes of the pseudo-observations, model simulations, and predictable component of the model simulations, respectively, so that their variance is 1; α_o and α_m are the signal-to-noise ratios (SNRs) of pseudo-observations and model simulations, respectively; and β is the ratio of systematic model error within the predictable component of the model simulations. With these definitions we explicitly choose an error which scales with the signal (through coefficient α_0 or α_m). This allows for a direct test on the importance of the SNR. Note that the addition of an arbitrary model error will have no impact on our analyses and results. The arbitrary error can be decomposed into a component scaling with the signal and a constant systematic bias. The latter has no net effect because of the linearity of (2) and of the measures of prediction skill (which only deal with anomaly/variations, as described below).

To the contrary, it is worthwhile acknowledging that this argument does not hold for typical short-term predictions which deal with variations in which the signal has a nonzero mean. For such predictions, the analysis presented here would have to be adjusted.

This stochastic model closely follows the one suggested by Siegert et al. (2016) with a small but crucial modification. Here the difference between observations and model is explicitly incorporated into the predictable component of the model rather than implicitly incorporated into the unpredictable component (i.e., noise) of the observations. Hence, in our analysis the observations are split in two terms representing their predictable and unpredictable components by nature; whereas in Siegert et al. (2016), observations are split in two terms representing components predictable and unpredictable by the model. It means that the Predictable Component defined by Eade et al. (2014) or Siegert et al. (2016) is actually the predicted component (i.e., what is predictable by the model), which is by construction smaller or equal to the Predictable Component by nature. (Indeed, what is predictable by the model is smaller or equal to what is predictable by nature.) This modification that we introduce to our stochastic model allows us for a more fundamental approach independent of model skills (i.e., independent of the model ability to accurately predict the full predictable component).

We use two diagnostics to measure the accuracy of the prediction: the skill and the reliability. The former is measured through the Coefficient of Determination or R^2 score. In the context of predicting the observations these read

$$R^2_{\text{obs}} = 1 - \frac{\overline{\left(\overline{m_i(t)}^i - o(t)\right)^2}^t}{\overline{o(t)^2}^t}, \tag{3a}$$

$$\text{Reliability}_{\text{obs}} = \sqrt{\overline{\left[\frac{\left(\overline{m_i(t)}^i - o(t)\right)^2}{\overline{\left(\overline{m_i(t)}^i - m_i(t)\right)^2}^i}\right]}^t}, \tag{3b}$$

where the overline corresponds to an average over either time (with the t superscript) or ensemble members (with i superscript). To obtain robust statistics we use 50,000 iterations for the time (t) together with a model ensemble of 4,000 members (i).

The Coefficient of Determination measures the similarity between the pseudo-observations and the model outputs (i.e., the predicted component). Multiplied by 100, it indicates the percentage of variance of the observations explained by the prediction. This means that a Coefficient of Determination of 1 indicates a perfect prediction, whereas a value of 0 indicates no prediction skill, equivalently a value of 0.5 indicates that 50% of the variance of the observations is represented by the prediction. On the other hand, the Reliability measures the consistency between the error and the model standard deviation (i.e., whether the unpredicted component is well captured by the ensemble member spread). Our formulation of Reliability follows the definition of Ho et al. (2013) but is generalized for nonstationary statistics following Sévellec and Drijfhout (2018). A value of 1 suggests a perfect consistency between the intrinsic prediction error (numerator) and the assessed prediction uncertainty (measured by the ensemble member spread, denominator). Values different from 1 indicate the unreliability of the prediction system. Hence, a value of 2 suggests that the prediction uncertainty is twice as small as the prediction error (corresponding to an overconfident prediction system), equivalently a value of 0.5 suggests that the prediction uncertainty is twice as big as the prediction error (corresponding to an underconfident prediction system).

To test the ability of the model to predict its own simulations (i.e., perfect model approach), these diagnostics are used after replacing the pseudo-observations by a certain model realization (a single member of the ensemble). The choice of the realization does not matter since all the model ensemble members have the same statistical behavior (and statistics have converged for our choice of time iterations).

2.2. Results

2.2.1. Impact of SNR on Prediction Accuracy

Using this statistical model, the prediction skills are diagnosed for a variety of SNR of both model and pseudo-observations (α_m and α_o, respectively). We first assume the absence of a systematic model error

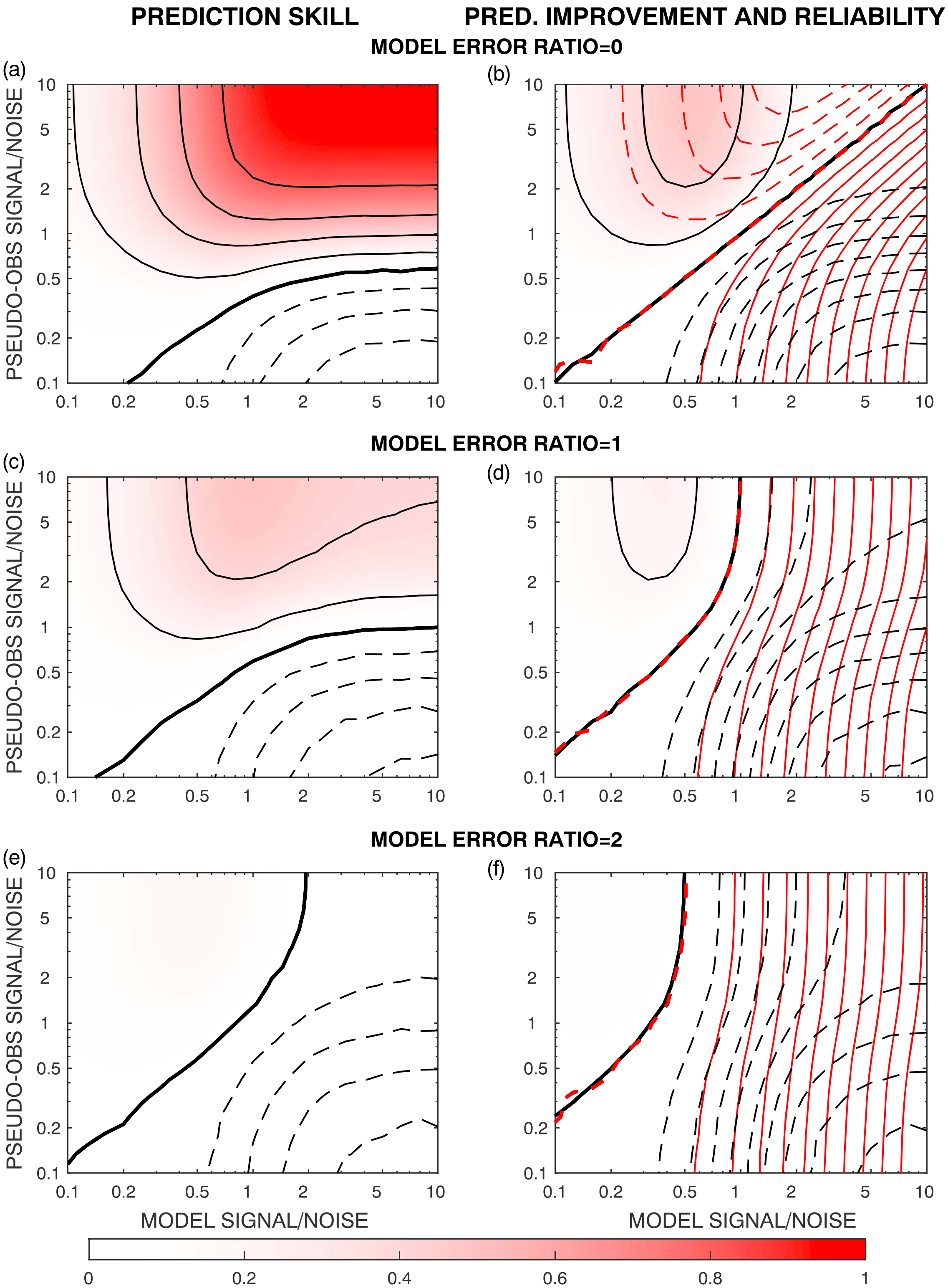

Figure 2. Prediction skill. (a, c, e) Coefficient of determination (the predicted component, R^2) for a pseudo-observation prediction following (3a) as a function of the predictable components of pseudo-observations and model (i.e., α_o and α_m, respectively), with systematic model error ratio (β) of 0, 1, and 2. (b, d, f) Skill difference between pseudo-observation prediction and perfect model approach, with systematic model error ratio (β) of 0, 1, and 2. Positive values show a parameter range leading to the signal-to-noise paradox (i.e., lower prediction skill in perfect model approach). The thick solid black lines show zero values, black thin solid and dashed lines show higher and lower values than zero, respectively, with contour intervals of 0.2. The thick dashed red lines show the reliability of 1, red thin solid and dashed lines show higher and lower reliability values than 1, respectively, with contour intervals of power of 10.

between model and pseudo-observations (i.e., $\beta = 0$), so they only differ through their relative level of predictable signal. In this case, the predictable and predicted components are equal, hence we recover the results from Siegert et al. (2016), which are summarized below for completeness.

Hence, within a perfect model approach, the skill, R^2, increases as a function of the model SNR through a $\alpha_m^2/(1+\alpha_m^2)$ law (Figure S1a), implying that increasing the relative amplitude of the predictable component leads to more skillful predictions. We also find that the reliability is always 1 (i.e., the model is perfectly reliable to predict itself, Figure S1b in the supporting information), showing that the model behavior is accurately sampled in a statistical sense.

For predicting pseudo-observations, the skill is always improved by increasing the pseudo-observation SNR, regardless of the model SNR (Figure 2a). However, the skill decreases when increasing the model SNR beyond that of the pseudo-observation SNR, which is clearly visible in case the pseudo-observation SNR < 1. In such cases, the prediction can have absolutely no skill ($R^2 < 0$).

The skill improvement when predicting pseudo-observations instead of the model itself can be expressed as the difference between the two coefficients of determinations ($R^2_{\mathrm{obs}} - R^2_{\mathrm{mod}}$, Figure 2b). An improvement of skill ($R^2_{\mathrm{obs}} > R^2_{\mathrm{mod}}$) always occurs when the pseudo-observation SNR is higher than the model SNR, as suggested by Eade et al. (2014), because the Ratio of Predictable Components (RPC) is larger than 1. This means that the signal-to-noise paradox is a natural outcome for models featuring a weaker signal in the prediction than is present in the observations, as long as model and (pseudo-)observations share the same predictable signal, that is, the model error is zero ($\beta = 0$).

To further understand the impact of the model SNR and the pseudo-observation SNR being different, we computed the prediction reliability (red contours in Figure 2b). An important property emerges from this diagnostic: a reliability of 1 can only be achieved if the RPC = 1. If the signal-to-noise paradox occurs (RPC > 1), the reliability of the prediction is decreased. In this case, the prediction of the ensemble spread is overdispersive (Figure 2b, Scaife et al., 2014; Siegert et al., 2016). This is a crucial result of the signal-to-noise paradox since, by measuring the plausibility and consistency of the prediction error, the reliability is arguably the most important property of a prediction system, in particular, if one wants to achieve probabilistic predictions and to provide risk assessments (Weisheimer & Palmer, 2014).

2.2.2. Impact of Systematic Model Error

These results change in the presence of a systematic model error (Figures 2c–2f). When $\beta \neq 0$, the predictable components (α_o and α_m for observations and model simulations, respectively) differ from the predicted components.

The role of systematic model error is illustrated by setting β in (2) to 1 and 2 (i.e., an error as big as and twice as big as the predictable component in model and pseudo-observations, respectively). For these two different levels of systematic model error, we find that the signal-to-noise paradox ($R^2_{\mathrm{obs}} > R^2_{\mathrm{mod}}$) is still possible (Figures 2d and 2f), but the regime of its occurrence becomes smaller for larger model error. Similar to the case without systematic model error, this occurs when the model SNR (α_m) is larger than the pseudo-observation SNR (α_o), but now with a threshold (upper bound), limiting the paradox to cases of low model SNR depending on the level of model error (Figures 2d and 2f). These upper bounds move to lower model SNR and larger RPC for increasing model error. This threshold/upper-bound breaks down the direct relation between the signal-to-noise paradox and the RPC. However, even in case of a strong model error (twice as big as the predictable component) a regime exists where the signal-to-noise paradox occurs. Since models always have some kind of systematic error (potentially significant), we can conclude that the signal-to-noise paradox is both a signature of a relatively too low model SNR (i.e., high RPC) and a signature that the model SNR is weak (i.e., <1).

The reliability of an erroneous model can still be 1, but this occurs only for RPC values larger than 1. In such cases, while the condition ($\alpha_o > \alpha_m$) applies, the prediction uncertainty and the model ensemble uncertainty can become statistically equivalent. Like in the case without systematic model error, the most accurate reliability is achieved when the signal-to-noise paradox is absent (Figures 2b, 2d, and 2f), regardless of the level of model error. We also find that, even under a significant level of systematic model error, the occurrence of the signal-to-noise paradox corresponds to an overdispersive regime in terms of ensemble spread prediction. Hence, the conclusion that the signal-to-noise paradox is the signature of an underconfident and unreliable prediction system is robust to the level of model error, even in cases of a large error of two (twice the predictable component) where the prediction skill is very weak (Figure 2e). It also appears that, in the presence of model error, a perfect RPC (= 1) corresponds to an underdispersive ensembles and so to a signature of an unreliable, overconfident prediction system.

3. Application to Climate Models

We now apply the framework of the signal-to-noise paradox to evaluate models from CMIP5 (5th Coupled Model Intercomparison Project) archive through their long forced historical simulations. Such simulations are not initialized with observations and not designed for interannual prediction. However, ideally they still feature similar statistical behavior as the observations. The question we want to answer is whether the signal-to-noise paradox occurs in models used in predictive systems. To this end, we compute the persistence of their annual-mean SAT between 1881 and 2004 for global and local spatial averages. The persistence is often used as the null-hypothesis of climate prediction and corresponds to assuming that the temperature will not change. Hence, the rate of persistence is an underestimation of the predicted component, and we will assume that the former can be used to approximate the latter to diagnose the signal-to-noise paradox (as suggested by Strommen & Palmer, 2018). In reality, predictability and the role of the SNR in this will depend on the state of the system. To investigate this in detail, one has to address initialized predictions, which is beyond the scope of the present study. Here, our main focus is an illustration of the concept developed above in coupled climate models. However, it is worth noting that the signal-to-noise paradox has been shown for global mean SAT with two different state-of-the-art prediction systems (DePreSys and PROCAST from Eade et al., 2014 and Sévellec & Drijfhout, 2018, respectively). Hence, we apply our analysis of persistence to 1- to 5-year hindcast lags (beyond 5 years persistence in models and observations becomes unskillful).

As described in the previous section, the predicted component or prediction skill is estimated by the coefficient of determination (R^2), reformulated for real observations and CMIP5 models. Hence, the global and local coefficients of determination for persistence read

$$R^2_{\text{Global}}(\tau) = 1 - \frac{\overline{[\text{GMT}(t+\tau) - \text{GMT}(t)]^2}^t}{\overline{\left[\text{GMT}(t) - \overline{\text{GMT}(t)}^t\right]^2}^t}, \tag{4a}$$

$$R^2_{\text{Local}}(x, y, \tau) = 1 - \frac{\overline{[\text{SAT}(x, y, t+\tau) - \text{SAT}(x, y, t)]^2}^t}{\overline{\left[\text{SAT}(x, y, t) - \overline{\text{SAT}(x, y, t)}^t\right]^2}^t}, \tag{4b}$$

where GMT is the Global Mean Temperature corresponding to the global spatial averaged SAT, τ is the hindcast lag set from 0 to 5 years with 1 year timesteps, x and y are the zonal and meridional coordinates, t is time going from 1881 to 2004. Thus, these two formulae are applied to three sets of observations and all members (with a minimum of three) of the nine climate models tested (see Appendix A: Method for details). In particular, applying the diagnostic to the full range of model-ensemble members and a range of observational reconstructions allows us to estimate the robustness of the results.

Focusing on the global scale and using (4a), we find that for GMT on average observations are more persistent than models (Figure 3a), with a good consistency between the three sets of observations. However, the model-observation difference becomes only significant after 2 years of prediction. Looking in more detail at individual models (Figures 3b and 3c), it appears that in the three models (CCSM4, IPSL-CM5A-LR, and MPI-ESM-LR), and to a lesser degree (FIO-ESM), persistence of GMT is comparable to observations. This leaves five other models with a too weak persistence in GMT, suggesting these models exhibit the signal-to-noise paradox for globally averaged SAT. This means that prediction systems for GMT based on these models are potentially underconfident and so unreliable. This also implies a GMT spectrum that is less red in these five models than in the observations. However, our analysis in general reveals a rather good agreement between observations and climate models (Figure 3a, with a relative error of 34% on average), suggesting a rather weak signal-to-noise paradox for GMT-predictions.

To characterize the signal-to-noise paradox on local scales we computed the skill of local persistence following (4b) for $\tau = 1$, 2, and 5 years (Figure 4 and S2). Because the local persistence for 2 and 5 years is extremely weak, we mainly concentrate on $\tau = 1$ year. The results of the local coefficients of determination (R^2_{Local}) are summarized by two indices. The first one is the Level of Agreement and measures for each

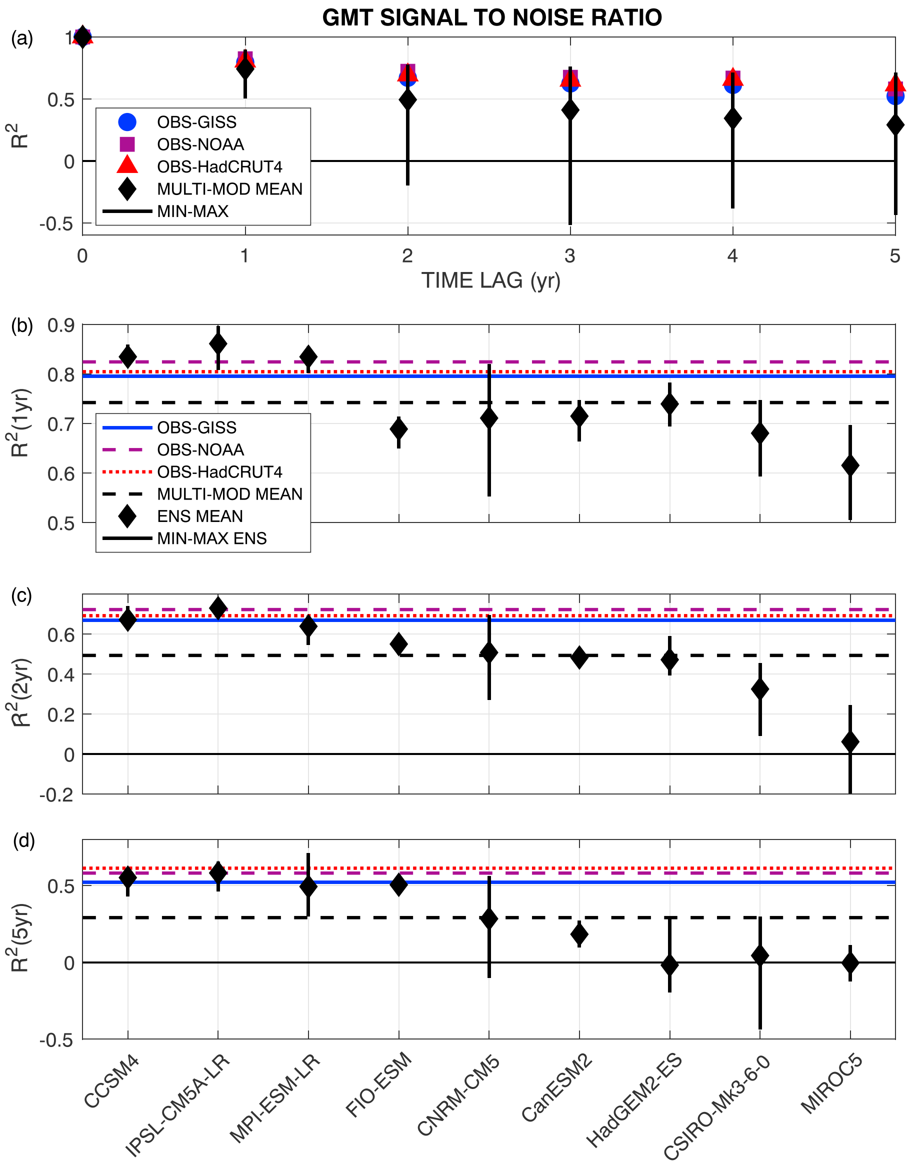

Figure 3. Global mean surface atmospheric temperature persistence in observations and CMIP5 models. (a) Persistence hindcast skills, measured through the coefficient of determination (R^2), averaged from 1 to 5-year hindcast lags. Hindcasts are annual averages from 1881 to 2004. Individual model values are shown for (b) 1-year, (c) 2-year, and (d) 5-year hindcast lags. Models with a lower prediction skill than observations are prone to exhibit the signal-to-noise paradox. Models have been sorted from the closest to observations to farther away, over the 5-year time lags tested.

climate model the relative area of the globe (in %) that has a local R^2 within ±10% of each of the three observational ones. The second index is the level of paradox and measures for each climate model the relative area of the globe (in %) that has a local R^2 smaller than each of the three observational ones, suggesting the occurrence of the signal-to-noise paradox.

The analysis shows a Level of Agreement that is extremely low, with a value below 10% for all nine models (Figure 4m). Analyses made with each of the three different observational data sets show excellent consistency. This suggests that climate models do not represent accurately the observational SAT persistence at local scales (grid box size). Consistently, the Level of Paradox is extremely high (Figure 4n), with 60% to 80% of the globe exhibiting a weaker persistence in climate models than in observations (i.e., featuring the signal-to-noise paradox). Again, this analysis is consistent between the three observational data sets. Hence, the high level of paradox suggests that prediction systems based on these models would be mainly under-

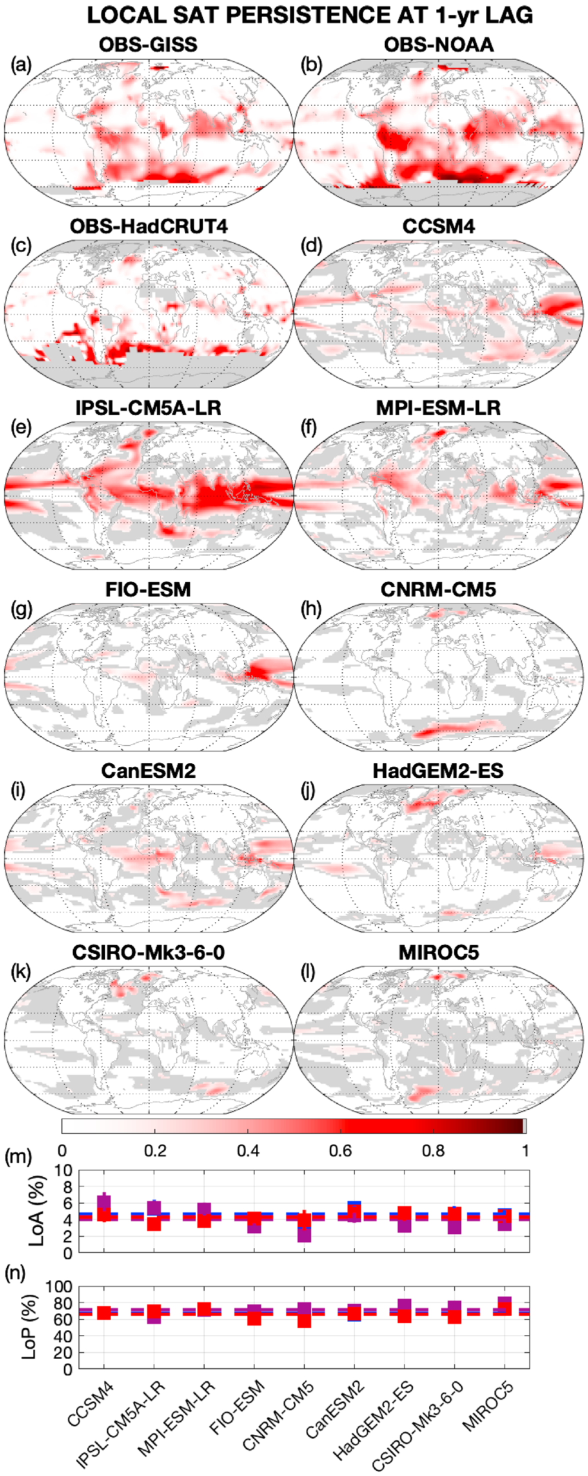

Figure 4. Local mean surface atmospheric temperature persistence in observations and CMIP5 models. Hindcasts are annually averages from 1881 to 2004 with a lag of 1 year. Surface Atmospheric Temperature persistence hindcast skills in (a–c) three observation data sets and (d–l) nine CMIP5 climate models (as the ensemble mean of member persistence). Gray indicates insufficient data for observations and a higher standard deviation than mean for models. (m) Level of Agreement (LoA) measures the relative area of the globe for which the models reproduces the observation persistence skills within a relative error of ±10%. (n) Level of Paradox (LoP) measures the relative area of the globe where the paradox occurs (i.e., weaker persistence skills in models than in observations). Red, purple, and blue indicate model comparisons with the GISS, NOAA, and HadCRUT4 observation data sets, respectively. Models have been sorted from the closest to observations to further away, over the 5-year time lags tested for GMT (following Figure 3).

confident and so unreliable. The 1-year persistence in observations (Figures 4a–4c) is mainly localized over the ocean. This suggests that the signal-to-noise paradox takes its source from the ocean rather than from land where the spectrum is less red. This means that models are relatively too noisy over the ocean (not enough variance on annual and longer time scales relative to shorter time scales) compared to observations over the ocean and feature a too white spectrum in SAT. Because longer time scales in SAT over the ocean dominantly arise from ocean variability, this suggests that ocean variability or ocean-to-atmosphere forcing are too weak in models. On longer time scales, the absence of persistence in both observations and models mechanically increases the level of agreement (models and observations having no-skill in more regions). This leads to the trivial solution that with weaker prediction skill the level of paradox is also weaker (Figure S3). However, there is still a nonnegligible level of paradox of between 30% and 70% of the globe at longer time scales.

4. Summary and Conclusions

In this analysis, we first tested the signal-to-noise paradox through a simple stochastic statistical representation of observations and model simulations. In particular, we set the statistical properties of pseudo-observations and model outputs to be virtually indistinguishable. The analysis confirmed that, in the absence of model error, the signal-to-noise paradox occurs when the relative part of the predictable component is weaker in the model than in observations. However, when systematic error is explicitly acknowledged, this relation breaks down. Indeed, a perfect Ratio of Predictable Components ($\mathrm{RPC} = 1$) leads to an unreliable, overconfident prediction system. It also restricts the signal-to-noise paradox to low model SNR. Hence, the weaker predictable component in models relative to observations becomes a necessary condition for the signal-to-noise paradox to occur but is no longer a sufficient condition anymore. Indeed, the necessary and sufficient condition becomes that the relative part of the predicted component is weaker in the model than in observations.

Hence, by adapting the RPC, we introduce the Ratio of Predicted Components: $\mathrm{R\Pi C} = \Pi\mathrm{C}_{\mathrm{obs}}/\Pi\mathrm{C}_{\mathrm{mod}}$ ($= R_{\mathrm{obs}}/R_{\mathrm{mod}}$), where $\Pi\mathrm{C}_{\mathrm{obs}}$ and $\Pi\mathrm{C}_{\mathrm{mod}}$ are the predicted components in observations and in the model, respectively. This leads to the trivial result, since it is its definition, that $\mathrm{R\Pi C} > 1$, and so $\Pi\mathrm{C}_{\mathrm{obs}} > \Pi\mathrm{C}_{\mathrm{mod}}$, is a signature of the signal-to-noise paradox. Since $\Pi\mathrm{C}_{\mathrm{mod}} = \mathrm{PC}_{\mathrm{mod}}$(i.e., a model can estimate its own predictability) and $\Pi\mathrm{C}_{\mathrm{obs}} \leq \mathrm{PC}_{\mathrm{obs}}$ (i.e., models underestimate the actual predictability of the observations), we have $\mathrm{RPC} \geq \mathrm{R\Pi C}$, so that RPC overestimates the actual occurrence of the signal-to-noise paradox. (Note that in the absence of systematic model error the predictable component and predicted component are strictly identical.) The accuracy of RΠC over RPC, demonstrates that the signal-to-noise paradox is a consequence of a given prediction system, rather than a fundamental property of the observations and their predictability. Our analysis also confirms the result from Eade et al. (2014) that the signal-to-noise paradox is a signature of an underconfident (overdispersive) prediction system. Hence, predictions can still be accurate but need a large number of members for the noisy unpredictable component to average out (Kumar, 2009). This also means that the signal-to-noise paradox is a sign of the weak reliability of a prediction system and could be used to estimate the system reliability.

Applying this new definition to previous work of Eade et al. (2014) on the North Atlantic Oscillation with a seasonal prediction system (GloSea5), we have $\Pi\mathrm{C}_{\mathrm{obs}} = 0.6$ and $\mathrm{R\Pi C}=\Pi\mathrm{C}_{\mathrm{obs}}/\Pi\mathrm{C}_{\mathrm{mod}} = 2.3$. This leads to $\mathrm{PC}_{\mathrm{mod}} = \Pi\mathrm{C}_{\mathrm{mod}} = 0.26$ and to $\alpha_m = \sqrt{R^2_{\mathrm{mod}}/(1 - R^2_{\mathrm{mod}})} = 0.27$. However it remains impossible to estimate $\mathrm{PC}_{\mathrm{obs}}$, but we know $\mathrm{PC}_{\mathrm{obs}} \geq 0.6$ (because $\mathrm{PC}_{\mathrm{obs}} \geq \Pi\mathrm{C}_{\mathrm{obs}}$), which leads to $\alpha_o=\sqrt{R^2_{\mathrm{obs}}/(1 - R^2_{\mathrm{obs}})} \geq 0.75$. From our analysis (Figure 2) this regime leads to the paradox for all tested errors and the conclusion that GloSea5 is an underconfident prediction system (consistently with Eade et al., 2014; Scaife et al., 2014).

Despite this nice result, it is however important to note that our stochastic model and the subsequent analysis of persistence only deal with anomalies. Hence, more common interannual prediction methods, which deal with short-term variations in which the signal has a nonzero mean, have also other sources of error, which were not considered here (e.g., systematic bias). To acknowledge these other types of error, the stochastic model presented here will have to be adjusted. This will be part of a follow-up study.

To diagnose the signal-to-noise paradox in state-of-the-art climate models, we computed the SAT persistence in nine climate models from CMIP5 and compared it to the SAT persistence in three observational data sets. We find that CMIP5 models have an important tendency to underestimate SAT persistence, conducive to the signal-to-noise paradox. This is particularly true over oceanic regions and at smaller spatial scales. This low level of persistence suggests that models can be improved by improving (enhancing) SAT-variance at longer time scales, especially over the ocean. Such improvement would most likely also lead to more reliable forecasts for the associated prediction systems. In light of this CMIP5 model analysis, investigating the SNR through initialized predictions in a range of state-of-the-art prediction systems would be a sensible and worthwhile effort.

The weaker persistence in models compared to the observations can be due to inaccurate observations showing too much persistence. Indeed, the three sets of observations tested are reconstructed from sparse and irregular in situ and remote observations. The reconstruction methods, which heavily depend on large spatial and temporal scale correlations to fill gaps and to extrapolate/interpolate missing data, can lead to overweighting slow variability and persistence in reconstructed observational products. This should be investigated in the future.

Another hypothesis to explain the signal-to-noise paradox, especially visible over the ocean, is the lack of SAT persistence in climate models. The ocean dynamics and ocean-atmosphere coupling are the most likely sources of this lack of persistence and should be targeted to improve the agreement between climate models and observations. To test the robustness of our analysis regarding the lack of persistence in SAT, we computed the 10 main empirical orthogonal functions (EOFs) of SAT in observations and in CMIP5 models. This analysis reveals that the models with more variance explained by their 10 main EOFs are less inclined to exhibit signal-to-noise paradox (Figure S4), especially when focusing on longer time scales. This suggests that the signal-to-noise paradox may come from the lack of coherent large-scale modes of SAT variability in the models and from the models featuring too much small-scale variability. A too weak ocean feedback onto the atmosphere has been found before (e.g., Haarsma et al., 2016) and since this aspect improves significantly in higher resolution models (Foussard et al., 2019; Minobe et al., 2008; Su et al., 2018), the next generation of climate models and their associated predictions system may suffer less from the signal-to-noise paradox, becoming more reliable, and more useful for operational probabilistic forecasts. However, it should be emphasized that the signal-to-noise paradox is not only due to a lack of model resolution and also points to more fundamental shortcomings in terms of physical processes incomplete or inadequately represented in those models.

Appendix A: Method

The model SAT was estimated from nine CMIP5 historical simulations restricted from 1881 to 2004 (Taylor et al., 2012). The nine models, with their number of members used in square brackets, are as follows: “CCSM4” [6], “CNRM-CM5” [10], “CSIRO-Mk3-6-0” [10], “CanESM2” [5], “HadGEM2-ES” [5], “IPSL-CM5A-LR” [6], “FIO-ESM” [3], “MPI-ESM-LR” [3], and “MIROC5” [5]. These models have been selected from the CMIP5 database because they have at least three members and the required data fields. We have set three ensemble members as the minimum for obtaining model result uncertainties and so to test their robustness. For observations, the GISS, NOAA, and HadCRUT4 temperature data sets were used.

Acknowledgments
This research was supported by the Natural and Environmental Research Council UK (SMURPHS, NE/N005767/1) and by the DECLIC and Meso-Var-Clim projects funded through the French CNRS/INSU/LEFE program. The authors acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output (listed in the Appendix of this paper). For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

Boer, G. J., Kharin, V. V., & Merryfield, W. J. (2019). Differences in potential and actual skill in a decadal prediction experiment. *Climate Dynamics*, *52*, 6619–6631.

Clark, R. T., Bett, P. E., Thonton, H. E., & Scaife, A. A. (2017). Skilful seasonal predictions for the European energy industry. *Environmental Research Letters*, *12*(119), 602.

Dunstone, N. J., Smith, D. M., Scaife, A., Hermanson, L., Eade, R., Robinson, N., et al. (2016). Multi-year predictability of the tropical Atlantic atmosphere driven by the high latitude North Atlantic Ocean. *Nature Geoscience*, *9*, 809–814.

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstonbe, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*, 5620–5628. https://doi.org/10.1002/2014GL061146

Foussard, A., Lapeyre, G., & Plougonven, R. (2019). Storm tracks response to oceanic eddies in idealized atmospheric simulations. *Journal Climate*, *32*, 445–463.

Haarsma, R. J., Selten, F. M., & Drijfhout, S. S. (2016). Decelerating Atlantic meridional overturning circulation main cause of future west European summer atmospheric circulation changes. *Environmental Research Letters*, *10*, 94007.

Ho, C. K., Hawkins, E., Shaffrey, L., Bröcker, J., Hemanson, L., Murphy, J. M., et al. (2013). Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion. *Geophysical Research Letters*, *40*, 5770–5775. https://doi.org/10.1002/2013GL057630

Kumar, A. (2009). Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Monthly Weather Review*, *137*, 2622–2631.

Kumar, A., Peng, P., & Chen, M. (2014). Is there a relationship between potential and actual skill? *Monthly Weather Review*, *142*, 2220–2227.

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., et al. (2015). Decadal climate prediction: An update from the trenches. *Bulletin of the American Meteorological Society*, *95*, 243–267.

Minobe, S., Kuwano-Yoshida, A., Komori, N., Xie, S.-P., & Small, R. J. (2008). Influence of the Gulf Stream on the troposphere. *Nature*, *452*, 206–209.

Palin, E. J., Scaife, A. A., Wallace, E., Pope, E. C. D., Arribas, A., & Brookshaw, A. (2016). Skillful seasonal forecasts of winter disruption to the u.K.transport system. *Journal of Applied Meteorology and Climatology*, *55*, 325–344.

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R., Dunstone, R. N., et al. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, *41*, 2514–2519. https://doi.org/10.1002/2014GL059637

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

Sévellec, F., & Drijfhout, S. S. (2018). A novel probabilistic forecast system predicting anomalously warm 2018–2022 reinforcing the long-term global warming trend. *Natures Communications*, *3024*, 9.

Siegert, S., Stephenson, D. B., Sansom, P. G., Scaife, A. A., Eade, R., & Arribas, A. (2016). A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *Journal of Climate*, *29*, 995–1012.

Smith, D. M., Cusack, S., Colman, A. W., Folland, C. K., Harris, G. R., & Murphy, J. M. (2007). Improved surface temperature prediction for the coming decade from a global climate model. *Science*, *317*, 796–799.

Smith, D. M., Eade, R., Dunstone, N. J., Fereday, D., Murphy, J. M., Pohlmann, H., & Scaife, A. A. (2010). Skilful multi-year predictions of Atlantic hurricane frequency. *Nature of Geosciences*, *3*, 846–849.

Strommen, K., & Palmer, T. N. (2018). Signal and noise in regime systems: A hypothesis on the predictability of the North Atlantic Oscillation. *Quarterly Journal of the Royal Meteorological Society*, *45*, 147–163.

Su, Z., Wang, J., Klein, P., Thompson, A. F, & Menemenlis, D. (2018). Ocean submesoscales as a key component of the global heat budget. *Nature Communications*, *9*, 775.

Svensson, C., Brookshaw, A., Scaife, A. A., Bell, V. A., Mackay, J. D., Jackson, C. R., et al. (2015). Long-range forecasts of UK winter hydrology. *Environmental Research Letters*, *10*, 64006.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 485–498.

Weisheimer, A., Decremer, D., MacLeod, D., O'Reilly, C., Stokdale, T. N., Johnson, S., & Palmer, T. N. (2019). How confident are predictability estimates of the winter North Atlantic Oscillation? *Quarterly Journal of the Royal Meteorological Society*, 1–20. https://doi.org/10.1002/qj.3446

Weisheimer, A., & Palmer, T. N. (2014). On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface*, *11*(20131), 162.