Received: 18 November 2017 | Revised: 21 February 2018 | Accepted: 22 February 2018 | Published on: 2 April 2018

DOI: 10.1002/asl.815

Atmospheric Science Letters RMetS

RESEARCH ARTICLE

Flow dependent ensemble spread in seasonal forecasts of the boreal winter extratropics

Dave MacLeod[1] | Chris O'Reilly[1] | Tim Palmer[1] | Antje Weisheimer[2,3]

[1]Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK

[2]Department of Physics, National Centre for Atmospheric Science (NCAS), University of Oxford, Oxford, UK

[3]Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

Correspondence
D. MacLeod, Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK.
Email: macleod@atm.ox.ac.uk

Funding information
FP7 Environment, Grant/Award number: 308378; Natural Environment Research Council, Grant/Award number: NE/MM005887/1; NERC SummerTIME, Grant/Award number: NE/MM005887/1; EU-funded EUCLEIA, Grant/Award number: 607085; EU-FP7 SPECS, Grant/Award number: 308378; NCAS Tim Palmer 2017/2018.

Flow-dependent spread (FDS) is a desirable characteristic of probabilistic forecasts; ensemble spread should represent the expected forecast error. However this is difficult to estimate for seasonal hindcasts as they tend to have a relatively small sample size. Here we use a long (110 year) seasonal hindcast dataset to evaluate FDS in forecasts of boreal winter North Atlantic Oscillation (NAO) and Pacific North American pattern (PNA). A good FDS relationship is found for interannual variations in both the NAO and PNA, with mild underdispersion for negative NAO and PNA events and slight overdispersion for positive NAO. Decadal-scale variability is seen in forecast errors but not in ensemble spread, which shows little variation on this timescale. Links between forecast errors and tropical heating anomalies are also investigated, though no strong links are found. However, a weak link between strong El Niño warming in the East Pacific and reduced PNA error is suggested.

KEYWORDS
boreal winter, ensemble spread, NAO, PNA, seasonal prediction, uncertainty

1 | INTRODUCTION

Ideally the spread generated by an ensemble forecast is a good indication of its error. This can be assessed by confronting multiple ensemble forecasts with the corresponding observations and calculating the average error of all forecasts with an ensemble spread within a certain range. For useful flow-dependent spread (FDS) we require a positive correlation between the error and spread: an ensemble with a larger spread should have a larger error, on average. FDS has been demonstrated in forecasts at multiple timescales (e.g., Ferranti *et al.,* 2015) by comparing historical reforecasts (hindcasts) with observed data.

Shorter-range forecasts generally have a large number of start dates, while seasonal forecasts initialized have significantly less, normally one per month. With around 30 years, a typical seasonal hindcast has a much smaller total number of reforecasts available to carry out forecast verification. This makes it harder to estimate FDS as reforecasts must be divided into bins, leading to small bin populations with large uncertainty in the average error calculated across their members. The verification of the spread-error relationship in seasonal forecasting systems is then generally limited to looking at the ratio between average spread and error or similar measures. This indicates whether on average the ensemble is under- or overdispersive, however, does not give an indication of FDS, that is, whether differences in the ensemble spread between individual forecasts is meaningful.

One approach to deal with the problem of small sample size in seasonal hindcasts is to take all spatial gridpoints within a region, under the assumption they are independent samples. Such an approach is used to generate reliability diagrams and has also been followed to verify FDS of both sea surface temperature (SST) in the European Centre For Medium-Range Weather Forecasts (ECMWF) seasonal

forecast system (Christensen *et al.,* 2015) as well as geopotential height at 500 hPa and temperature at 850 hPa in the monthly system (Haiden *et al.,* 2016). However when the target field is smooth (as upper-level atmospheric fields such as Z500 are) neighboring gridpoints are not independent. This method is also not appropriate for an index which does not have spatial dimensions.

A sample of years larger than the usual 30 is needed to estimate FDS of seasonal forecasts, and such a hindcast has recently been produced with the ECMWF atmospheric model. This is an atmosphere-only seasonal forecast covering the entire 20th century (hereafter ASF-20C, Weisheimer *et al.,* 2017) and with 110 years it is several times longer than other operational hindcasts. Here it is used to make a robust estimate of FDS, focusing on the main modes of atmospheric variability in the northern extratropics, the North Atlantic Oscillation (NAO) and Pacific North American pattern (PNA). Given the research interest in seasonal forecasts of the northern extratropics and their potential utility in decision-making (Scaife *et al.,* 2014; Shi *et al.,* 2015; Dunstone *et al.,* 2016; Weisheimer *et al.,* 2017), the question of whether they exhibit FDS is pertinent.

Using the large sample of the ASF-20C hindcast, here we assess FDS of the NAO and PNA and thus establish the information content of variations in forecast spread. Dependence of this relationship on the underlying mode of variability is also considered, as well as the long term variability in dispersion characteristics. Finally potential relationships between forecast error and atmospheric heating in the tropics are investigated. The methodology, results and discussion follow in subsequent sections.

2 | METHODOLOGY

2.1 | ASF-20C: Seasonal forecasts of the 20th century

The ASF-20C hindcast uses the ECMWF integrated forecasting system (IFS) atmosphere model, in seasonal forecast mode, but without a dynamic ocean. The ERA-20C reanalysis (Poli *et al.,* 2016) which assimilates only surface pressure and marine wind observations is used for initializing the model, as well as for validation of the hindcast, while SSTs from the HadISST2.1.0.0 dataset (Rayner *et al.,* 2003) are used to initialize and force the lower boundary. By forcing with observed SST rather than a dynamic ocean this hindcast provides an estimation of forecast skill assuming SST predictions are perfect. The hindcast is initialized on the first of every November across the period 1900–2009 and 4-month 51 member forecasts made for each year. The simulations are then averaged over lead month 2–4 in order to assess forecasts of the boreal winter season of DJF. Further details of the model setup and analysis of the hindcast may be found in Weisheimer *et al.* (2017) and O'Reilly *et al.* (2017).

2.2 | Measuring FDS in modes of atmospheric variability in the northern extratropical circulation

The NAO is the dominant mode of variability over the Euro-Atlantic sector and has a strong impact on the weather and climate over Europe. It shows variability on multiple timescales, from days to years and decades and forecasting it on seasonal timescales is a longstanding research challenge (Müller *et al.,* 2005; Palmer *et al.,* 2005; Scaife *et al.,* 2014; Weisheimer *et al.,* 2017). The PNA is a dominant teleconnection pattern in the extratropics during the Northern Hemisphere winter, strongly associated with regional temperature and precipitation anomalies across North America (Wallace and Gutzler, 1981). Indices for both the NAO and PNA are defined following Weisheimer *et al.* (2017) and Wallace and Gutzler (1981), respectively; readers are referred here for details. The indices are calculated in the same way for ERA-20C and ASF-20C and then normalized in both cases by the standard deviation of the ERA-20C index.

The FDS is then calculated for the 110 independent probabilistic forecasts of DJF NAO and PNA. The first step is to separate the years into bins according to their ensemble spread (defined as the standard deviation across the ensemble) and calculate the average error within each bin. Quartile boundaries of spread are used to define the bins, that is, the reforecasts are divided into four equally populated subsets according to their value of ensemble spread. For each bin the root mean square error (RMSE) between the ensemble mean and the reference is calculated and compared with the average ensemble spread, for each of the four bins. In addition the slope of the regression line between the four binned values of RMSE and spread is calculated to quantify the relationship between spread and error. Note that an individual forecast may have a large ensemble spread but a small ensemble mean error. It is quite possible that an observation may fall close to the ensemble mean of ensemble with high spread; only on average we expect the ensemble mean RMSE to be larger when the ensemble spread is larger. For this reason the average RMSE is calculated across multiple reforecasts with similar spread.

Four bins were chosen as a compromise between robustly estimating the average RMSE in each bin by maximizing bin population when the data is subset by positive and negative phases of the NAO and PNA and highly resolving the relationship between RMSE and spread by maximizing the total number of bins. For each bin a 95% confidence interval for the average error is estimated from 1,000 resamples with replacement from that bin. The 1,000 resamples are also used to estimate the uncertainty in the calculated correlation between the RMSE and spread, according to one standard error (one standard deviation of 1,000 correlations from the resamples). For the analysis of links between tropical SST and forecast error five (quintile) bins are used instead of four, as no subsetting by phase is

required. Bootstrap resampling is also used to calculate the significance of SST anomalies; years are randomly resampled from the entire 110 year dataset 1,000 times and a 95% confidence interval centered on a zero anomaly is calculated; SST anomalies outside of this interval are then considered to be significant at the 95% level.

3 | RESULTS

3.1 | FDS of the NAO and PNA

Figure 1a,d shows the FDS relationship for DJF NAO and PNA forecasts across all 110 years. The slope across all points is within one standard error of unity for both indices, indicating that in general the spread is a good indication of the error. For the NAO the forecast spread varies between 0.8 and 1.1 of the NAO index, indicating the possible range of ensemble spread which the prediction system forecasts. For the PNA the forecast spread is smaller with slightly narrower range, around 0.6–1.0 of the PNA index. The PNA exhibits some nonlinearity in the relationship, such that the forecasts in the lowest spread bin do not necessarily have the lowest error and so small differences between low spread forecasts may not be meaningful. The average spread-error measured across the entire dataset (blue star) indicates that spread is roughly equal to the error, on average.

Figure 1b,c shows the same relationship for separate phases of the NAO, defined according to the reference data. For NAO-events the error over all points is larger than the spread and the gradient of the regression is more than one standard error above unity, though the confidence interval for all bins encompasses the diagonal. The sample size precludes a strong statement about the difference between NAO+ and NAO− events, though suggests slight underdispersion (overdispersion) for NAO− (NAO+) events. For the PNA a similar difference is observed, with PNA-events showing potential underdispersion for the highest spread events, though again this is not significant. For the second quintile, there appears to be significant overdispersion for PNA-events.

Considering now low-frequency variability in dispersion characteristics, Figure 2a shows the spread, RMSE and their ratio calculated for 30 year running windows for the NAO across the 110 year period. RMSE has larger variability

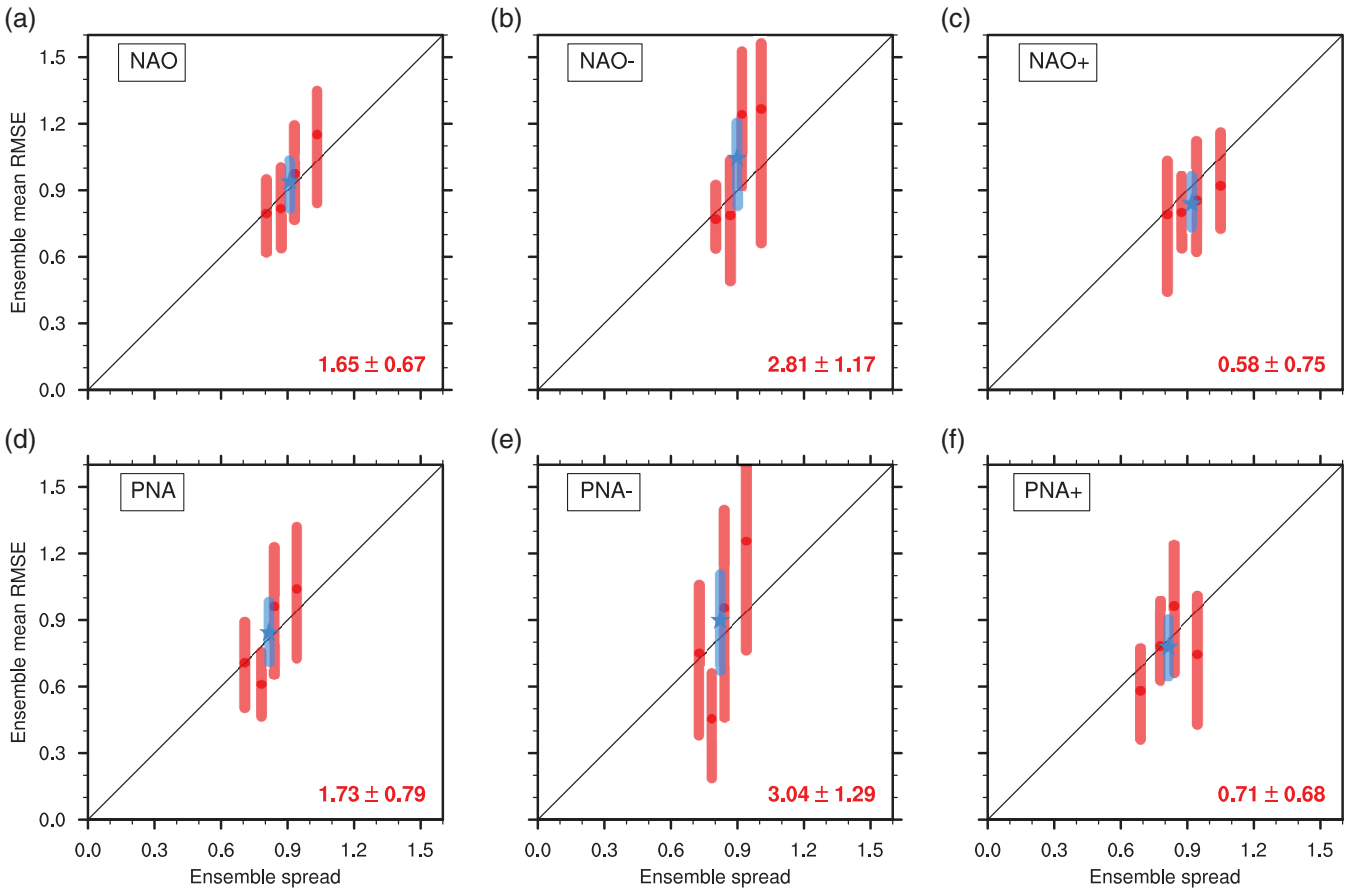

FIGURE 1 Flow dependent spread (FDS) of the DJF forecasts for the (a–c) NAO and (d-f) PNA. Forecasts are binned into equally-populated bins of ensemble spread and the average RMSE in each bin is plotted as a small circle. Error bars indicate 95% confidence intervals from 1,000 member bootstrap resampling within each bin. The star indicates the spread-error estimate across the entire dataset. The black lines demarcate where the error is equal to the spread and the lower-right number represents the slope of the regression line across the points, along with the standard error. (a) and (b) show the FDS over all NAO and PNA years, while (b, c) and (e, f) show the same for just NAO+/− and PNA+/− years, defined according to the reference data

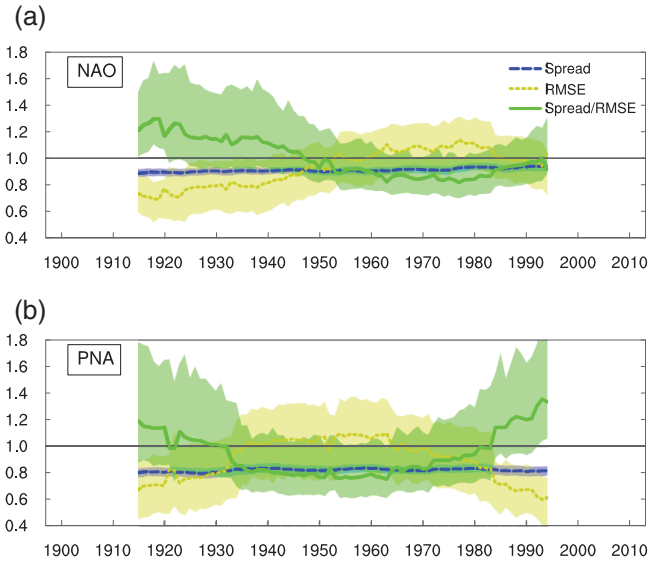

FIGURE 2 30 year average spread, RMSE and the ratio between them for (a) the NAO and (b) the PNA. Shaded areas indicate 95% confidence intervals from 1,000 member bootstrap resampling

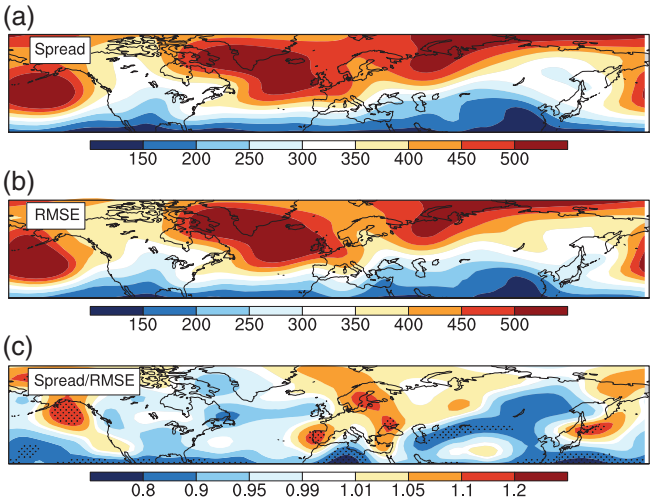

FIGURE 3 Maps of (a) spread, (b) RMSE and (c) the ratio between them for DJF anomalies of Z500. The units for the spread and RMSE plots are meters. Stippling for spread/error comes from a 1,000 member resampling, indicating deviation of the ratio from 1 at 95% significance

than spread, which exhibits no variability on this timescale. RMSE is lower in the first half of the century, higher in the mid to late century and has a slight negative trend in the recent period. This is consistent with the analysis of Weisheimer *et al.* (2017) which finds higher NAO skill in earlier and later periods. Since the ensemble spread shows no variability, the spread/RMSE ratio follows the inverse of RMSE; it is overdispersive earlier and underdispersive in the later period. However the 95% confidence intervals of spread and error overlap throughout most of the period suggesting that for most of the hindcast period the spread is roughly consistent with the error.

The PNA also shows no low-frequency variability in ensemble spread (Figure 2b). The RMSE is significantly below (above) one at the beginning and end of the period (during the middle period). This is consistent with O'Reilly *et al.* (2017) where a mid-century drop in PNA predictability is found. The dispersion ratio then follows the inverse of the RMSE, with overdispersion at the start and end of the period and underdispersion during the middle, though the 95% confidence interval includes unity except for at the end of the period. Overall Figure 2 suggests slight deviations from a well-calibrated spread/RMSE forecast on decadal timescales, though these deviations are not highly significant. It is clear that the spread does not show variability on long timescales, and so variability spread/RMSE arises from RMSE variability alone. If there is some process leading to decadal variations in RMSE, this does not seem to be represented in the hindcasts.

In order to evaluate the spatial characteristics of dispersion, the spread and RMSE in DJF Z500 across the entire Northern Hemisphere is shown in Figure 3. Overall the pattern of spread strongly resembles the pattern of RMSE. There is a gradient in RMSE from the tropics to the midlatitudes due arising from lower variability of Z500 closer to the equator. This gradient is well captured by the ensemble spread, though there is significant underdispersion in the lower part of the domain extending southward to the tropics. Over the United States and Atlantic the spread/RMSE ratio is not different from unity at 95% confidence, while a large coherent region in East Asia is underdispersive. There are some regions of overdispersion over Southern Europe, as well as in the West and Northeast Pacific.

3.2 | Tropical drivers of forecast uncertainty

The potential relationship between SST forcing and NAO/PNA error was investigated. Figure 4 shows SST anomalies associated with the years in each spread quintile. For the NAO (Figure 4a–e) warm anomalies in the Central and West Pacific and the Indian Ocean are associated with the forecasts with the highest spread. The SST pattern associated with the lowest spread forecasts is associated with cold SST in the Central Pacific. Given the linear relationship between forecast spread and error (Figure 1), this suggests that warm anomalies here are related to uncertainty in the forecast. When the sea surface is warm, there is a tendency for the forecast to have a larger error; conversely when the Central Pacific is cool, the error on average may be smaller. The same analysis was reproduced for NAO negative and positive years separately but no clear differences were found (not shown).

For the PNA (Figure 4f–j) there is warming in the East Pacific for the subset of forecasts with the lowest spread, suggesting that El Niño may act to drive the model toward slightly more predictable state in the extratropics. There is no clear SST pattern associated with larger forecast spread. Stratifying this analysis by PNA positive and negative years reveals that the relationship between low ensemble

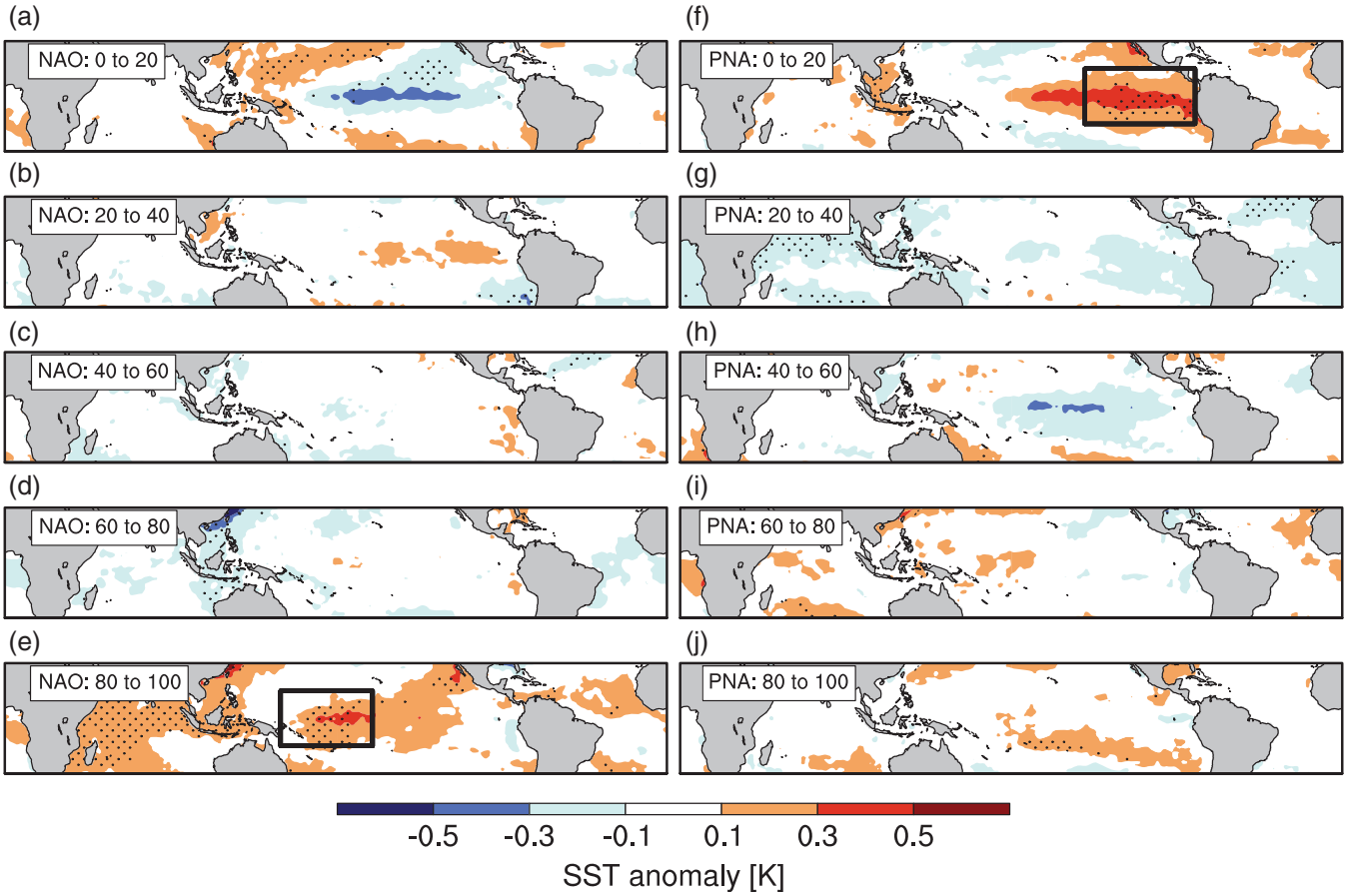

FIGURE 4 Average DJF SST in the tropics, stratified by the ensemble spread in the (a–e) NAO and (f–j) PNA. The hindcast data is stratified according to equally spaced bins, and is arranged from the years with the lowest spread, (a) and (f) to the years with the highest spread. Stippling indicates differences significant from zero at the 95% confidence level

spread and East Pacific warming was arising PNA+ years alone (not shown). This is somehwat expected, due to the known relationship between the PNA and El Niño (Straus and Shukla, 2002). To investigate further, we define a potential forcing box in the tropics for each of the NAO and PNA according to Figure 4. For the NAO we define a Central Pacific region and for the PNA an East Pacific region (shown in Figure 4e,f, respectively) and plot the absolute error for each index against the average SST in the box.

Figure 5a shows the relationship between NAO error and SST in the Central Pacific, while Figure 5b shows the same for the PNA and SST in the East Pacific. For the NAO a few of the warmest years have large errors, but there are insufficient data points showing the same relationship to draw any conclusions; over most of the distribution there does not appear to be any strong relationship. For the PNA, Figure 5b reveals that only the two very warmest years out of the total 110 are contributing to the association of warm SSTs with low spread seen in Figure 4f.

To further test for any link between tropical SSTs in these regions and atmospheric error, we consider the same analysis for in a coupled hindcast following the same setup as ASF-20C, coupled to NEMO. Hindcast data used has 25 member ensembles covering 1981–2014 and unlike ASF-20C, this hindcast includes initial perturbations to the atmospheric state, but otherwise the setup is the same. Here, each of the 25 members has a diverging SST field for each of 34 years, giving a total sample size of 850 members over which to consider the relationship between SST and atmospheric error.

Results are shown in Figure 5c,d. Similarly to the uncoupled run there is no systematic relationship for either the NAO or PNA, demonstrating that variability in errors in the extratropical atmospheric seasonal forecast is not clearly related to variations in tropical heating. However, in the coupled system the 2 years with extreme heating in the East Pacific also lead to reduced PNA error across each ensemble, suggesting a link between strong El Niño events and low PNA error. However, with only a few years falling in the extreme of the distribution it is difficult to do more than speculate on the existence of such a link.

4 | DISCUSSION AND CONCLUSIONS

Using a long seasonal hindcast we have shown that the ensemble spread of forecasts for both the NAO and PNA is a good indicator of forecast error. The NAO forecast spread

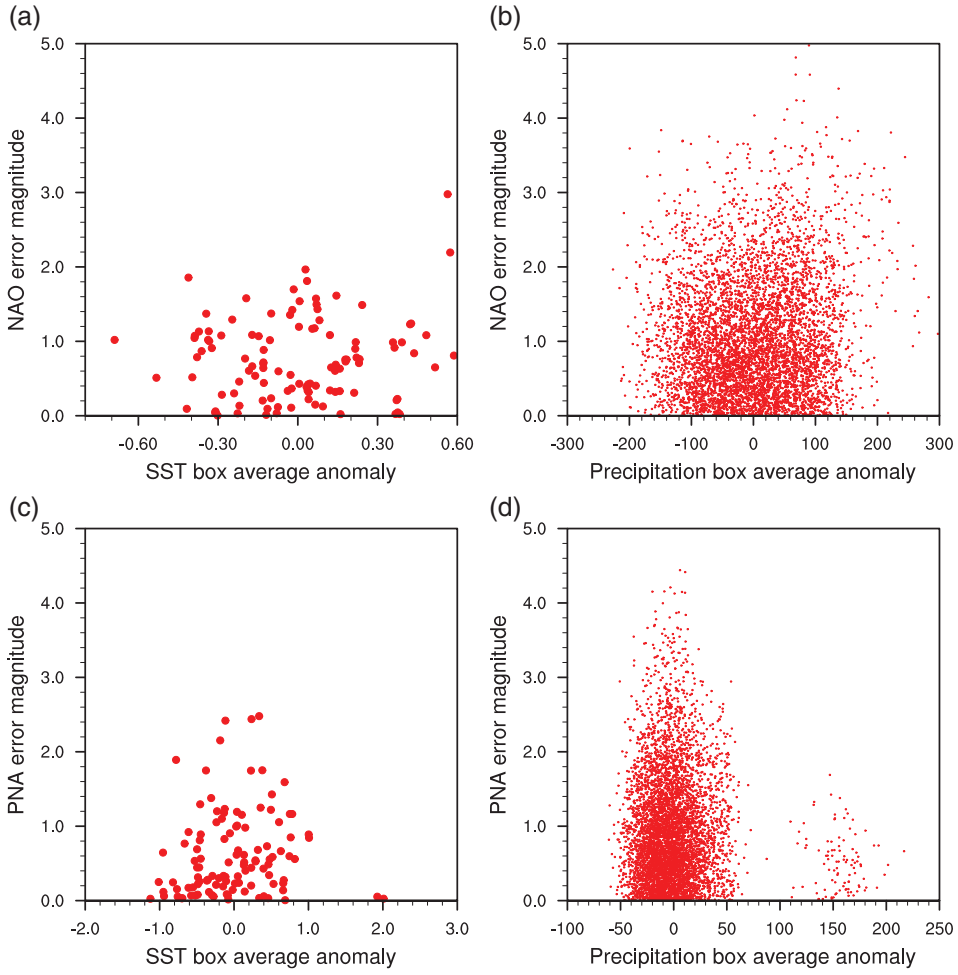

FIGURE 5 (a) DJF NAO ensemble mean RMSE against the DJF SST in the Central Pacific box corresponding to Figure 4e (b) as (a), for PNA and the East Pacific box corresponding to Figure 4f. (c,d) as (a,b) but for coupled seasonal hindcasts covering 1981–2014

varies between 0.8 and 1.1 of the NAO index while PNA forecast spread varies between 0.6 and 1.0 standard deviations of the PNA index and in each case the spread is positively correlated with the average error. These uncertainty forecasts have only slight resolution, that is, successive forecasts do not have a wide variation in their predicted uncertainty. This has implications for the utility of such forecasts. That is, the model forecast spread does not vary much from quite a wide uncertainty of the NAO/PNA index, and users should be aware of this. However this marginal information may be useful to some.

Some weak evidence is found for overdispersion for NAO- and PNA-events, as well as decadal-scale variability in dispersion characteristics, in agreement with previous studies using the same dataset (O'Reilly *et al.*, 2017; Weisheimer *et al.*, 2017). Observed decadal-scale variability in RMSE of both the PNA and NAO is not captured by hindcasts, which demonstrate no strong decadal-scale variability in spread. That is, long-term variability in dispersion characteristics is driven primarily by variations in error, not spread. Across the entire northern hemisphere extratropics, the pattern of spread matches error and their ratio is mostly indistinguishable from one, however significant areas of underdispersion are found for the tropics and East Asia. Significant overdispersion can be seen over smaller regions of the Pacific and southern Europe.

Potential links between tropical heating and extratropical atmospheric forecast error have also been explored, though no strong links are found. However, two warm outlying years of East Pacific heating are associated with low PNA forecast error, in both uncoupled and coupled simulations, suggesting that strong El Niño events lead to a more predictable, positive PNA. The current analysis only suggests this hypothesis and if it exists the effect is marginal. Future work might investigate this hypothesis through modeling experiments, for instance, testing to see if imposing warm anomalies in the East Pacific SSTs leads to reductions in PNA variability. If such an effect can be robustly demonstrated it may guide a priori evaluation of individual seasonal forecasts and improve their operational interpretation.

ACKNOWLEDGEMENTS

This work was supported in part by the EU-FP7 SPECS project (grant agreement number 308378), the EU-funded EUCLEIA project (grant agreement number 607085) the NERC SummerTIME project (NE/MM005887/1) and by the ERC Advanced Investigator Award 'Towards the

Prototype Probabilistic Earth-System Model'. The authors declare no conflict of interest.

ORCID

Dave MacLeod http://orcid.org/0000-0001-5504-6450

REFERENCES

Christensen, H.M., Moroz, I.M. and Palmer, T.N. (2015) Evaluation of ensemble forecast uncertainty using a new proper score: application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141, 538–549. https://doi.org/10.1002/qj.2375.

Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N., Andrews, M. and Knight, J. (2016) Skilful predictions of the winter North Atlantic Oscillation one year ahead. *Nature Geoscience*, 9, 809–814. https://doi.org/10.1038/ngeo2824.

Ferranti, L., Corti, S. and Janousek, M. (2015) Flow-dependent verification of the ECMWF ensemble over the euro-Atlantic sector. *Quarterly Journal of the Royal Meteorological Society*, 141, 916–924. https://doi.org/10.1002/qj.2411.

Haiden, T., Janousek, M., Bidlot, J., Ferranti, L., Prates, F., Vitart, F., Bauer, P. and Richardson, D.S. (2016) *Evaluation of ECMWF forecasts, including the 2016 resolution upgrade* (ECMWF Tech. Memo 792). Retrieved from http://www.ecmwf.int/sites/default/files/elibrary/2015/15275-evaluation-ecmwf-forecasts-including-2014-2015-upgrades.pdf

Müller, W.A., Appenzeller, C. and Schär, C. (2005) Probabilistic seasonal prediction of the winter North Atlantic Oscillation and its impact on near surface temperature. *Climate Dynamics*, 24, 213–226. https://doi.org/10.1007/s00382-004-0492-z.

O'Reilly, C.H., Heatley, J., MacLeod, D., Weisheimer, A., Palmer, T.N., Schaller, N. and Woollings, T. (2017) Variability in seasonal forecast skill of Northern Hemisphere winters over the twentieth century. *Geophysical Research Letters*, 44, 5729–5738. https://doi.org/10.1002/2017GL073736.

Palmer, T.N., Doblas-Reyes, F.J., Hagedorn, R. and Weisheimer, A. (2005) Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 360, 1991–1998. https://doi.org/10.1098/rstb.2005.1750.

Poli, P., Hersbach, H., Dee, D.P., Berrisford, P., Simmons, A.J., Vitart, F., Laloyaux, P., Tan, D.G.H., Peubey, C., Thépaut, J.-N., Trémolet, Y., Hólm, E.V., Bonavita, M., Isaksen, L. and Fisher, M. (2016) ERA-20C: an atmospheric reanalysis of the twentieth century. *Journal of Climate*, 29, 4083–4097. https://doi.org/10.1175/JCLI-D-15-0556.1.

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

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. (2014) Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, 41, 2514–2519. https://doi.org/10.1002/2014GL059637.

Shi, W., Schaller, N., MacLeod, D., Palmer, T.N. and Weisheimer, A. (2015) Impact of hindcast length on estimates of seasonal climate predictability. *Geophysical Research Letters*, 42, 1554–1559. https://doi.org/10.1002/2014GL062829.

Straus, D.M. and Shukla, J. (2002) Does ENSO force the PNA? *Journal of Climate*, 15, 2340–2358. https://doi.org/10.1175/1520-0442(2002)015⟨2340:DEFTP⟩2.0.CO;2.

Wallace, J.M. and Gutzler, D.S. (1981) Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review*, 109, 784–812. https://doi.org/10.1175/1520-0493{↖%}281981{↖%}29109{↖%}3C0784{↖%}3ATITGHF{↖%}3E2.0.CO{↖%}3B2.

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D.A. and Palmer, T. (2017) Atmospheric seasonal forecasts of the twentieth century: multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution. *Quarterly Journal of the Royal Meteorological Society*, 143, 917–926.

How to cite this article: MacLeod D, O'Reilly C, Palmer T, Weisheimer A. Flow dependent ensemble spread in seasonal forecasts of the boreal winter extratropics. *Atmos Sci Lett*. 2018;19:e815. https://doi.org/10.1002/asl.815