Geophysical Research Letters

RESEARCH LETTER
10.1029/2020GL090059

Representing Model Uncertainty in Multiannual Predictions

Daniel J. Befort[1], **Christopher H. O'Reilly**[1,2], **and Antje Weisheimer**[2,3]

[1]Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK, [2]National Centre for Atmospheric Science, University of Oxford, Oxford, UK, [3]European Centre for Medium-Range Weather Forecasts, Reading, UK

Key Points:

- SST predictions from single-model ensembles tend to be overconfident/underdispersive on multiannual time scales up to 28 months
- The effectiveness of stochastic single-model ensembles and multimodel combinations to improve forecast reliability has been studied
- Stochastic schemes as efficient, low-cost alternatives to represent model uncertainty should be used in conjunction with multimodels

Supporting Information:

- Supporting Information S1

Correspondence to:
D. J. Befort,
daniel.befort@physics.ox.ac.uk

Citation:
Befort, D. J., O'Reilly, C. H., & Weisheimer, A. (2021). Representing model uncertainty in multiannual predictions. *Geophysical Research Letters*, *48*, e2020GL090059. https://doi.org/10.1029/2020GL090059

Received 4 AUG 2020
Accepted 24 NOV 2020

Abstract The most prominent way to account for model uncertainty is through the pragmatic combination of simulations from individual climate models into a multimodel ensemble (MME). However, alternative approaches to represent intrinsic model errors within single-model ensembles (SMEs) using stochastic parameterizations have proven beneficial in numerical weather prediction. Nevertheless, stochastic parameterizations are not included in most current decadal prediction systems. Here, the effect of the stochastically perturbed physical tendency (SPPT) scheme is examined in 28-month predictions using ECMWF's forecast model and contrasted with a MME constructed from current decadal prediction systems. Compared to SMEs, SPPT improves the skill and reliability of tropical sea surface temperature forecasts during the first 18 months (similar to the MME). Thus, stochastic schemes can be an effective and low-cost alternative to be used separately or in conjunction with the multimodel combination to improve the reliability of climate predictions on multiannual time scales.

Plain Language Summary To obtain reliable predictions on any time scale, it is inevitable to account for model uncertainties caused by unresolved processes. One prominent way to do this is by combining simulations from different models into a multimodel ensemble (MME). However, in numerical weather prediction, it has been shown that using stochastic physics, which aims to represent the effect of the unresolved processes, is another possibility to account for model uncertainties within a single model. Here, we assess in how far stochastic physics improves skill and reliability of predictions on multiannual time scales up to 28 months. It is found that tropical sea surface temperatures tend to be overconfident (and thus unreliable) in single-model ensembles. Reliability can be largely increased not only by using a MME but also by using stochastic physics for forecast times up to about 18 months. This shows that stochastic schemes can be considered an effective and low-cost alternative to be used separately or in conjunction with the multimodel combination to improve the reliability of climate predictions on multiannual time scales.

1. Introduction

The inevitable approximations needed to solve the equations of the laws of physics in state-of-the-art climate models are a major source of error and uncertainty in model simulations of current and future climate. In general, circulation climate models sub-grid-scale tendencies of prognostic variables are represented, or parameterized, as functions of the resolved variables. However, such parameterizations may not be consistent with underlying scaling symmetries of the dynamical equations or with observations of power law structure in the real atmosphere. The effects of sub-grid-scale variability on the resolved larger scales make it impossible to represent all scales of motion and their interactions explicitly within weather and climate models, leading to uncertainty across a range of temporal and spatial scales (Palmer, 2001, 2012).

Over recent years, the multimodel ensemble (MME) has emerged as a pragmatic and much used approach for representing the effects of model uncertainties. In particular, it was demonstrated for seasonal forecasts that a fixed-size ensemble constructed from different models is more reliable than an ensemble of the same size from a single model (Palmer et al., 2004; Stockdale et al., 2009; Weisheimer et al., 2009). Similar results comparing MMEs versus single-model ensembles (SMEs) were recently found for climate predictions and projections on decadal time scales (Verfaillie et al., 2020).

MMEs are, however, limited by the number of available independent models, each of which having their own error characteristics. While sampling these individual model errors can overcome some of the problems, the multimodel approach does not touch upon the problems of the sub-grid-scale variability affecting resolved scales.

Recognizing the fundamental sub-grid-scale uncertainties, the numerical weather prediction community has been at the forefront of developing new stochastic approaches. These approaches rely on stochastic parameterization schemes where the underlying deterministic subgrid parameterizations are replaced by an inherently stochastic formulation, to explicitly account for model-related uncertainties. In particular, stochastically perturbed physical tendency (SPPT) schemes for the atmosphere are now used routinely for global and regional numerical weather prediction (e.g., Berner et al., 2017; Leutbecher et al., 2017; Lock et al., 2019; Pegion et al., 2016).

Stochastic physics schemes are also increasingly applied in dynamical seasonal predictions where they have been shown to significantly improve the forecasts in the tropics through their impact on deep convection. Of particular importance is their ability to substantially increase the ensemble spread of the underdispersive sea surface temperature (SST) ensembles in the tropical oceans leading to improved reliability of ENSO and associated global teleconnections (Batté & Doblas-Reyes, 2015; Doblas-Reyes et al., 2009; Weisheimer et al., 2014). Comparing the performance of stochastic parameterizations to the multimodel approach, Weisheimer et al. (2011) concluded that on monthly time scales the ensemble forecast system with stochastic parameterization provided the overall most skillful probabilistic forecasts for temperature and precipitation. On seasonal time scales, the results depended on the variable, with the multimodel outperforming the stochastic approach for near-surface temperature but not for precipitation. It is important to note that implementing stochastic parameterizations in an SME comes at minimal computational cost.

There is increasing demand for skillful and reliable information for the next 10 years, which motivated the development of decadal prediction systems. While stochastic parameterizations have now started being utilized in forced climate simulations using Earth-system models (Christensen et al., 2017; Davini et al., 2017; Palmer, 2019), initialized decadal climate predictions are predominantly performed without stochastic parameterizations. One exception is the work from Corti et al. (2012), who analyzed an initialized decadal hindcast conducted using ECMWF's coupled model and found good reliability over several regions. However, this study did not focus to assess the impact of stochastic physics on skill and reliability. In this study, we extend the analysis of model uncertainty beyond seasonal prediction to the multiannual time scale using targeted simulations conducted with ECMWF's coupled forecasting model. These results have important implications for communities working on multiyear to decadal predictions.

A description of the methods and data sets used can be found in Section 2; results demonstrating the impact of stochastic physics on the tropics and extratropics are detailed in Section 3. Finally, the main findings are summarized in Section 4.

2. Methods and Data

Two hindcast sets using ECMWF's coupled model CY46R1 have been conducted to test the impact of stochastic physics (SPPT) on multiyear time scales. The setup of both experiments is identical: (i) initialized each 1st November from 1981 to 2014, (ii) 10 ensemble members, (iii) 28-month forecasts, and (iv) atmospheric horizontal resolution Tco199 (approx. 50 km); 1° ocean resolution. The only difference between the runs is that the stochastic physics scheme SPPT has been switched off (*ECMWF-noSPPT*) in one experiment, whereas it is included in the other experiment (*ECMWF-SPPT*).

In the SPPT scheme, the summed tendencies of the prognostic variables temperature, wind, and humidity as passed on from the individual parameterization schemes are perturbed with a multiplicative univariate Gaussian noise term. Such a multiplicative approach recognizes the flow-dependent uncertainty that arises from within the individual parameterization schemes, while also aiming to keep the physical consisten-

cy between the individual parameterized tendencies. The applied perturbations vary smoothly following an order 1 autoregressive (AR1) process in space and time with three distinct spatiotemporal scales with characteristic lengths of 500, 1,000, and 2,000 km. The corresponding temporal scales (*e*-folding times) are 6 hours, 3 days, and 30 days. The shortest scale is connected with the largest amplitude of the perturbations, whereas the longest and slowest scale becomes active via small perturbations. The choice of the amplitude of the perturbations has been motivated by results from coarse-graining studies with cloud-resolving models. Stochastic parameterizations have been extensively documented by the numerical weather forecasting and seasonal prediction community (e.g., Buizza et al., 1999; Leutbecher et al., 2017; Lock et al., 2019; Palmer et al., 2009; Weisheimer et al., 2014).

The stochastic physics approach is compared to five different decadal prediction systems listed in Table S1: NCAR-DPLE (Yeager et al., 2018) and four dcppA CMIP6 systems: EC-Earth (Bilbao et al., 2020; Doblas-Reyes et al., 2018; Haarsma et al., 2020), MPI-ESM1-2-HR (Mauritsen et al., 2019; Müller et al., 2018; Pohlmann et al., 2019), MIROC6 (Kataoka et al., 2020), and HadGEM3-GC31-MM (Andrews et al., 2020; Williams et al., 2018). Note that while NCAR-DPLE had 40 members, here only 10 members are used to allow a fair comparison to the other ensembles. Even though more decadal predictions from CMIP6 are available, here hindcasts used are limited to those initialized in November to match the initialization month of both ECMWF hindcasts.

Skill is measured using the anomaly correlation coefficient (ACC) of the ensemble mean. Furthermore, we assess the reliability of the different predictions using the spread-over-error (SoE) relationship, which is defined as the ratio between the average ensemble spread and the root-mean-square-error (RMSE) of the ensemble mean. Generally speaking, a perfect reliable forecasting system is one where the verification is indistinguishable from the ensemble. It can be shown analytically that for a perfect ensemble the time-mean ensemble spread (standard deviation around the ensemble mean) should equal the time-mean RMSE of the ensemble mean forecast (Palmer et al., 2006). The spread–skill relationship is often used in numerical weather prediction to guide model development for the ensemble prediction system. The relationship implies that for perfectly reliable predictions, the SoE measure equals 1, within the sampling uncertainty. Values of SoE > 1 indicate an overdispersive (underconfident) prediction system, whereas values SoE < 1 indicate underdispersive (overconfident) ensemble predictions.

The skill of both ECMWF simulations is further compared to a 10-member MME consisting of members from NCAR-DPLE and the four decadal prediction systems from CMIP6. The skill for both ECMWF ensembles is assessed based on a 10,000 sample bootstrap over years. In contrast, for the MME, we randomly sample over years and members. The latter allows to also include the uncertainty of the specific ensemble members included in the MME. For global reliability maps, we use four categories: significantly overconfident (SoE < 1 and confidence $> 95\%$), overconfident (SoE < 1 but confidence $< 95\%$), underconfident (SoE > 1 but confidence $< 95\%$), and significantly underconfident (SoE > 1 and confidence $> 95\%$).

From previous studies on the impact of stochastic physics in seasonal predictions, it has become clear that the tropical Pacific is the region where the stochastic schemes have the largest and significant impact. As the ENSO regions are also an important source of global teleconnections, we focus the analysis in this study on the SSTs over the NINO3 region (150–90°W; 5°S–5°N) of the eastern tropical Pacific for forecast times up to 28 months. We also show global statistics for several seasons which are included in the 28 months. Besides analyzing the effect of SPPT in the tropics, where several studies have shown the positive impact of SPPT, we also investigate the impact in midlatitudes. Specifically, we assess the skill of the extratropical large-scale atmospheric circulation by analyzing predictions of sea-level pressure (SLP) anomalies, how this varies in ensembles with and without SPPT, and also in comparison with the MME. Furthermore, the impact of SPPT on skill and reliability of the North Pacific index (180–120°W; 30–65°N), which is strongly influenced by tropical SSTs (O'Reilly, 2018), is assessed.

All hindcasts have been corrected for lead-time-dependent biases. Here, anomalies are calculated using all initialization dates between 1981 until 2014. SLP and SST data from ERA5 reanalysis are used as reference (Hersbach et al., 2020). All data sets have been interpolated to a common 2.5° grid.

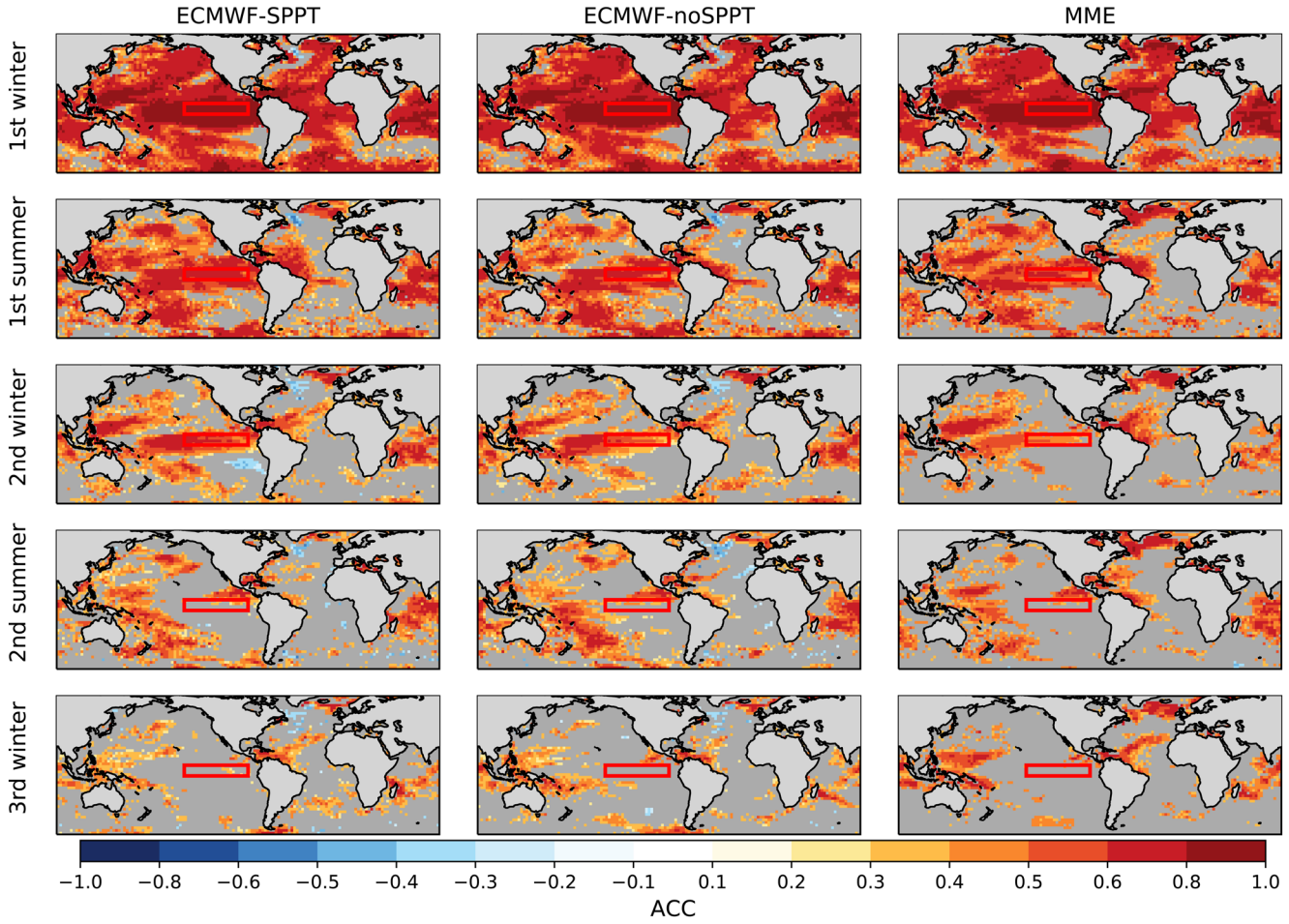

Figure 1. Anomaly correlation coefficient for SSTs and different forecast times (first to third winter/first and second summer). The 10-member MME median is calculated from a 10,000 bootstrap sample created by randomly selecting two different members from each single-model ensemble. Non-gray-shaded areas are significantly different to 0 with 95% confidence. Significance is calculated by sampling over years for *ECMWF-SPPT* and *ECMWF-noSPPT* ensembles and over years and members for the MME (10,000 samples). Red rectangle indicates the NINO3 region. SST, sea surface temperature; MME, multimodel ensemble; SPPT, stochastically perturbed physical tendency.

3. Results

3.1. Skill and Reliability of Global SSTs

ACC scores for global SSTs for the ECMWF ensemble *ECMWF-SPPT*, *ECMWF-noSPPT*, and the MME are shown in Figure 1. For the first winter (forecast months 2–4), significant positive skill is found over most parts of the globe and there are only minor differences between the three different ensembles. This is similar for SSTs during the first summer (forecast months 8–10) for which skill is found not only over tropical regions but also over large parts of the extratropics. Skill decreases further and during the second winter (forecast months 14–16) it is more restricted to the tropical regions, especially over the Pacific Ocean. Furthermore, larger differences between the three ensembles appear during second winter. Over the central Pacific, the MME shows smaller skill compared to both ECMWF ensembles, whereas the MME tends to be more skillful over the North Atlantic. The low skill in the ECMWF seasonal forecasting system over the North Atlantic has recently been attributed to the initialization of the Atlantic Meridional Overturning Circulation (Tietsche et al., 2020). After the second winter, skill decreases in all ensembles and is absent across most of the globe by third winter (forecast months 26–28), except over parts of the Warm-Pool region, the North Atlantic and the Indian Ocean.

Next, reliability of all ensembles is assessed using the SoE metric. In contrast to ACC skill, differences between the three ensembles are already apparent during the first winter (Figure 2). Both *ECMWF-SPPT* and

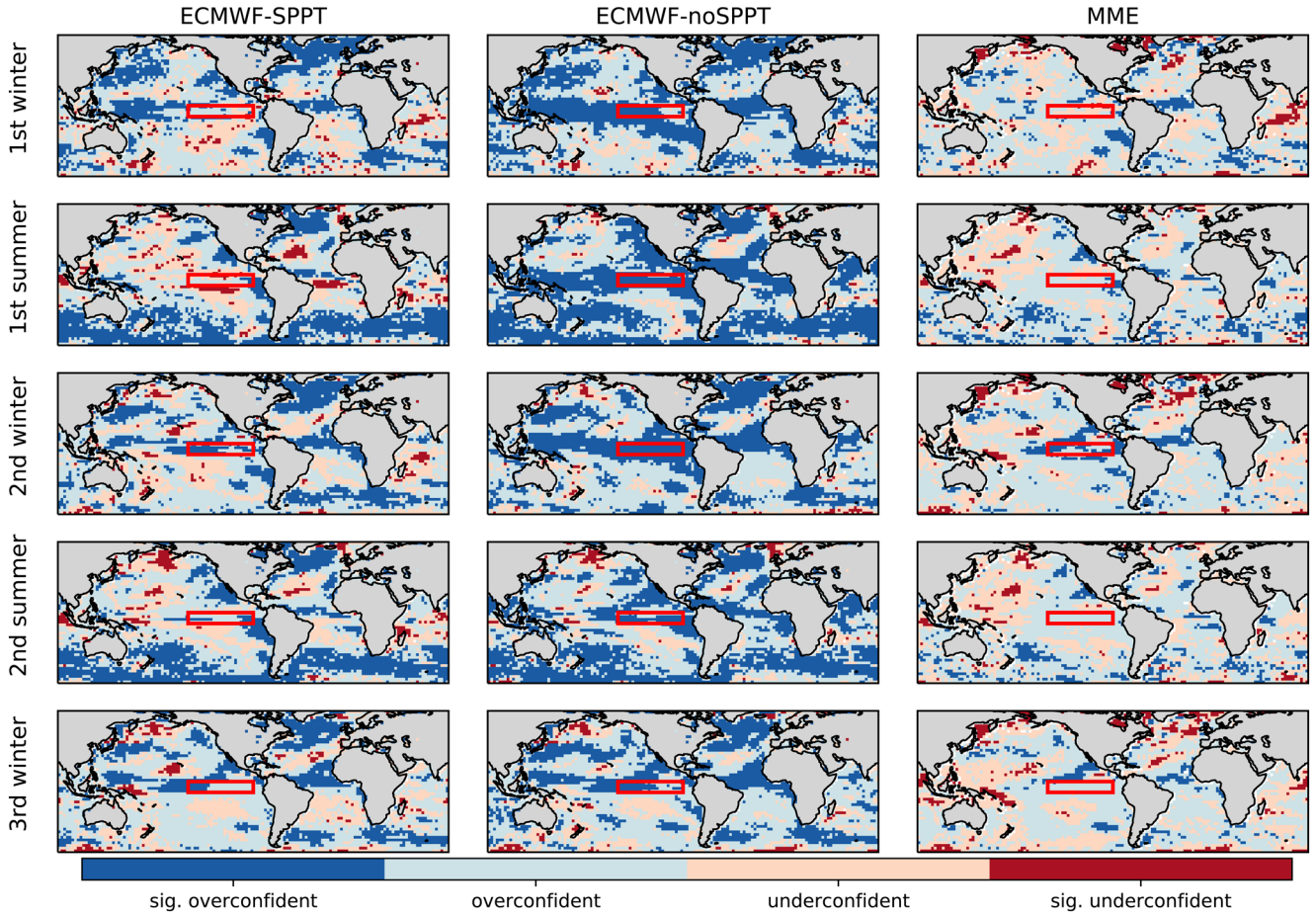

Figure 2. Same as Figure 1 but for spread-over-error (SoE). SoE values are categorized into significant overconfident, overconfident, underconfident, and significantly underconfident (see Section 2).

the MME are reliable over most parts of the globe and especially over the tropics, except over the tropical Atlantic. In contrast, *ECMWF-noSPPT* is significantly overconfident, which is especially pronounced over the tropical oceans. This improvement in reliability due to SPPT on seasonal time scales is in agreement with previous studies (Weisheimer et al., 2011). Until, the second summer reliability over the tropics is increased in the *ECMWF-SPPT* simulations compared to the *ECMWF-noSPPT* hindcast. For the latter, overconfidence over the tropical Pacific Ocean is apparent throughout the hindcast. In contrast the *ECMWF-SPPT* ensemble is reasonably reliable for the first and second winter as well as for the second summer.

Results for the MME indicate a reliable ensemble over most of the globe and especially over large areas of the tropical Pacific Ocean up to the third winter. These findings are in agreement with previous studies that have demonstrated the improved reliability of multimodel compared to SMEs (Doblas-Reyes et al., 2009; Palmer et al., 2004; Stockdale et al., 2009; Verfaillie et al., 2020; Weisheimer et al., 2009). To compare the level of reliability of all three ensembles quantitatively, we analyzed a number of grid cells which are significantly unreliable over the tropical band (20°N–20°S) (Table S2). It is found that for all lead times, the reliability of the MME is highest, which shows about half the number of significantly unreliable grid cells for each season compared with the *ECMWF-SPPT* hindcast. The SSTs in the *ECMWF-noSPPT* hindcast are the most unreliable with about twice the number of unreliable SST grid cells compared with the *ECMWF-SPPT* hindcast.

3.2. Skill and Reliability of ENSO Indices

To further investigate the impact of stochastic physics ensembles and MMEs on skill and reliability over the tropical Pacific, we now analyze the hindcast performance of ENSO indices.

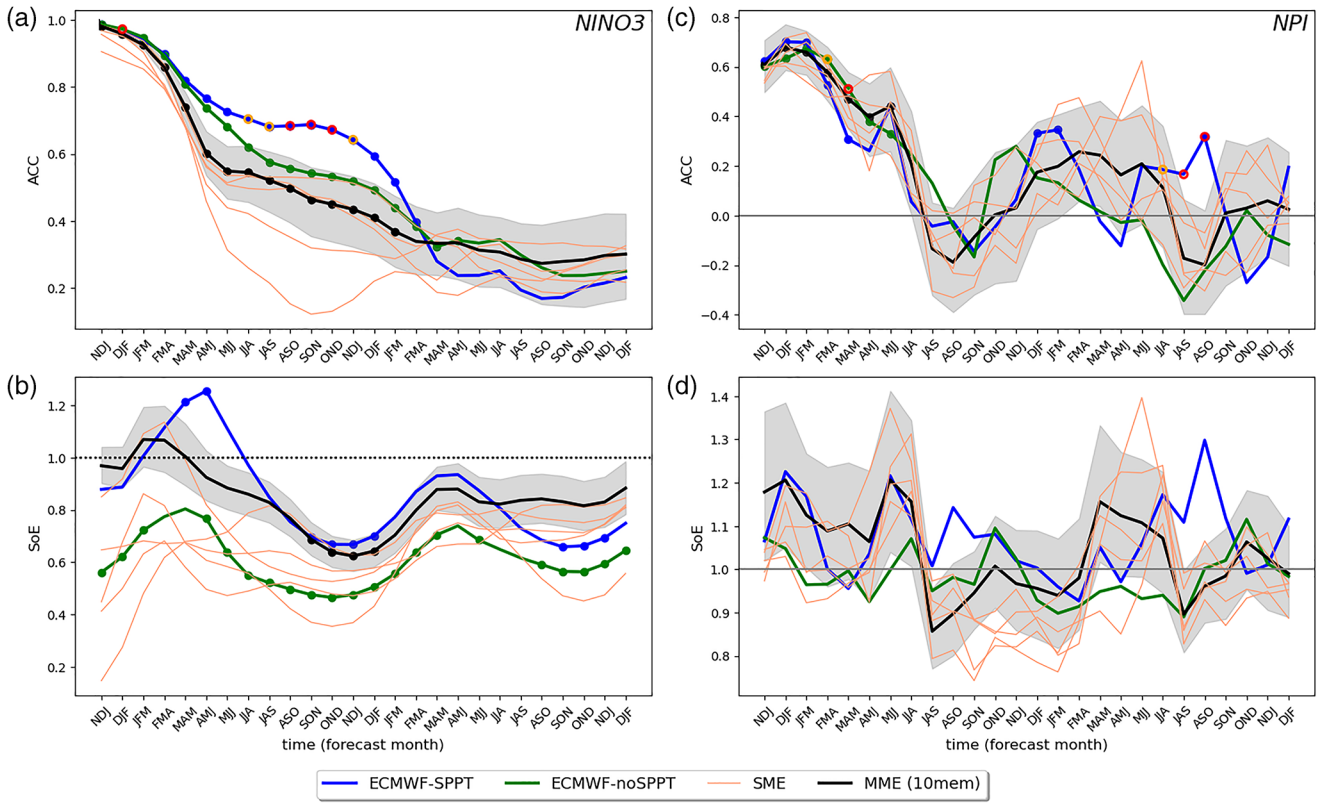

Figure 3. (a) Anomaly correlation coefficients for SSTs over NINO3 region using ERA5 as reference, (b) same as (a) but for SoE, (c) same as (a) but for the North Pacific index (NPI), (d) same as (c) but for SoE. Gray shading for the MME indicates 2.5 and 97.5 percentile derived from randomly sampling (10,000 samples) two members from each single-model ensemble. Dots in (a) and (c) indicate forecast times for which the respective ensemble is significantly larger than 0, whereas dots in (b) and (d) indicate forecast times for which the respective ensemble is significantly different from 1 (95% confidence, 10,000 samples). Samples have been generated by bootstrapping over years for *ECMWF-SPPT* and *ECMWF-noSPPT* ensembles and over years and members for the MME. Orange and red circles in (a) and (c) indicate those forecast times for which the respective ECMWF ensemble shows significantly higher skill compared to the other ECMWF ensemble (orange: 10% and red: 5% significance level, following Siegert et al., 2017). SST, sea surface temperature; SoE, spread-over-error; MME, multimodel ensemble; SPPT, stochastically perturbed physical tendency.

Figure 3a shows ACC skill for 3-month-averaged SSTs over the NINO3 tropical eastern Pacific region for each ensemble. All three simulations (MME, *ECMWF-SPPT*, and *ECMWF-noSPPT*) show significant positive correlation skill up to the second winter. ACC is higher for *ECMWF-SPPT* compared to *ECMWF-noSPPT* up to the second spring (forecasts times: 16–17 months). To test the significance of differences in correlation skill, we use the method of Siegert et al. (2017). For the NINO3 region, significant improvements with SPPT are found during the first autumn season (Figure 3). Over the western tropical Pacific, using SPPT demonstrates significant improvement up to the second winter, whereas *ECMWF-noSPPT* provides significantly more skillful SST predictions from the second summer onwards (Figure S2). However, it should be noted that both ECMWF hindcasts only consist of 10 ensemble members each. A larger ensemble is needed to provide a robust assessment of skill improvement related to stochastic physics since previous studies using larger ensembles showed the positive impact of stochastic physics on Pacific SSTs on seasonal time scales (Subramanian et al., 2017; Weisheimer et al., 2014). After the second year spring barrier to ENSO predictability (Cane, 1991; Webster & Yang, 1992) in March–April–May, the ACC becomes similarly low and is no longer significant for any ensemble. The MME and most of the single-model decadal prediction ensembles show a stronger reduction of skill after the first spring than the two ECMWF ensembles but especially the *ECMWF-SPPT* simulation. However, the skill of the MME is sensitive to the choice of ensemble members (indicated as gray shading in Figure 3a).

As already described in Section 3.1, both MME and *ECMWF-SPPT* ensembles exhibit higher reliability than the *ECMWF-noSPPT* ensemble, which is especially pronounced over the tropics. Figure 3b illustrates

reliability for the NINO3 region measured by the SoE statistic. It is found that reliability over this region is increased over the 28 months in ECMWF IFS model when using SPPT compared to not using SPPT. The *ECMWF-noSPPT* experiment is significantly unreliable (overconfident) for most forecast times, except during second spring to second summer, whereas ECMWF *ECMWF-SPPT* ensemble is only significantly unreliable during the first spring, first and second autumns and 2nd winter. The improvements due to stochastic physics are to a large extent related to increased spread within the ensemble but also due to a reduced error (Figure S3). SoE exhibits an annual cycle with lower values in autumn and higher values in spring in both ensembles, which is primarily related to an annual cycle in the RMSE. Furthermore, during the first half year of the forecasts, *ECMWF-SPPT* has an excessively large spread which is related to an SPPT modification introduced into the recent model cycle (for details see Lock et al., 2019). ECMWF's operational seasonal forecasts from SEAS5, based on an older model cycle, do not show such an overdispersion (Johnson et al., 2019). Comparing the different SMEs reveals that besides having different magnitude in error and spread, they are all overconfident for most forecast times. This is similar to what is found for the *ECMWF-noSPPT* ensemble, suggesting that it is a common feature of SMEs. While this has been known for seasonal time scales, it is shown here for the first time that the overconfidence continues beyond annual time scales. As suggested by previous studies, reliability can be increased by using a MME, particularly when the constituent models are themselves overconfident (Weigel et al., 2008). For the NINO3 region, we find that the MME is reliable for all forecast times up to 28 months except during the first autumn to second winter.

Similar results are found for the central Pacific NINO3.4 region (170–120°W; 5°N–5°S), whereas reliability is superior over the western central Pacific NINO4 region (160° E–150°W; 5°N–5°S) (see Figures S1 & S2).

3.3. Analysis of Atmospheric Circulation Skill in the NH Extratropics

Given the improvement seen in the tropical Pacific SSTs, a natural extension is to examine the influence of SPPT on extratropical circulation in the multiyear predictions. It is certainly plausible that the improved NINO3 predictions might translate into improved predictions of teleconnections to the extratropical circulation. We begin by analyzing the predictions of SLP in the NH extratropics. Maps of correlation skill for the *ECMWF-SPPT*, *ECMWF-noSPPT*, and the decadal MME are shown in Figure 4.

High levels of skill are evident over the tropics during the first winter, which is typical of seasonal forecasting systems (e.g., Smith et al., 2012). In the extratropics, significant levels of skill are generally limited to the extratropical North Pacific where ENSO teleconnections are strong (Trenberth et al., 1998), whereas elsewhere the signal due to the relatively small ensemble size becomes weak. Skill drops off in both the tropics and extratropics by the second winter across all ensembles but is most notable in the *ECMWF-noSPPT* ensemble and the MME. In the second winter, there remains some significant skill in the tropics in both the ECMWF ensembles and also, to a lesser extent, in the MME. However, the *ECMWF-SPPT* ensemble is the only system with substantial levels of skill in the extratropical North Pacific.

Essentially, no significant skill remains by the second summer, even in the tropics, consistent with the SST analysis in Sections 3.1 and 3.2. There is a curious feature in the third winter, however, consisting of a region of relatively high—and seemingly significant—skill in the extratropical North Pacific in the *ECMWF-SPPT* ensemble, as well as some significant skill over the western North Atlantic. Skill in this region is typically thought to originate from teleconnection with the tropical Pacific and is associated with skill in the tropical SLP (e.g., as seen in the first and second winters). SLP skill in the tropics is largely absent in the third winter in the *ECMWF-SPPT* ensemble, so it is not clear that the extratropical skill in the third winter is robust.

We now examine the evolution of the skill of the extratropical North Pacific SLP anomalies over lead time by analyzing a North Pacific index region (as shown by the blue boxes in Figure 4). The evolution of the correlation skill of the North Pacific SLP (NPI) index is shown in Figure 3c with significant skill indicated by the solid coloured dots in the plot. Across the extended first winter season, both the ECMWF IFS ensembles and the MME exhibit similarly high levels of skill. Into the spring season, skill in the *ECMWF-SPPT* ensemble drops off slightly more quickly than the *ECMWF-noSPPT* ensemble and by the first summer season none of the ensembles exhibit significant skill. Beyond the first summer season, neither the *ECMWF-noSPPT* ensemble nor the MME have significant correlation skill for the North Pacific SLP index. However, in

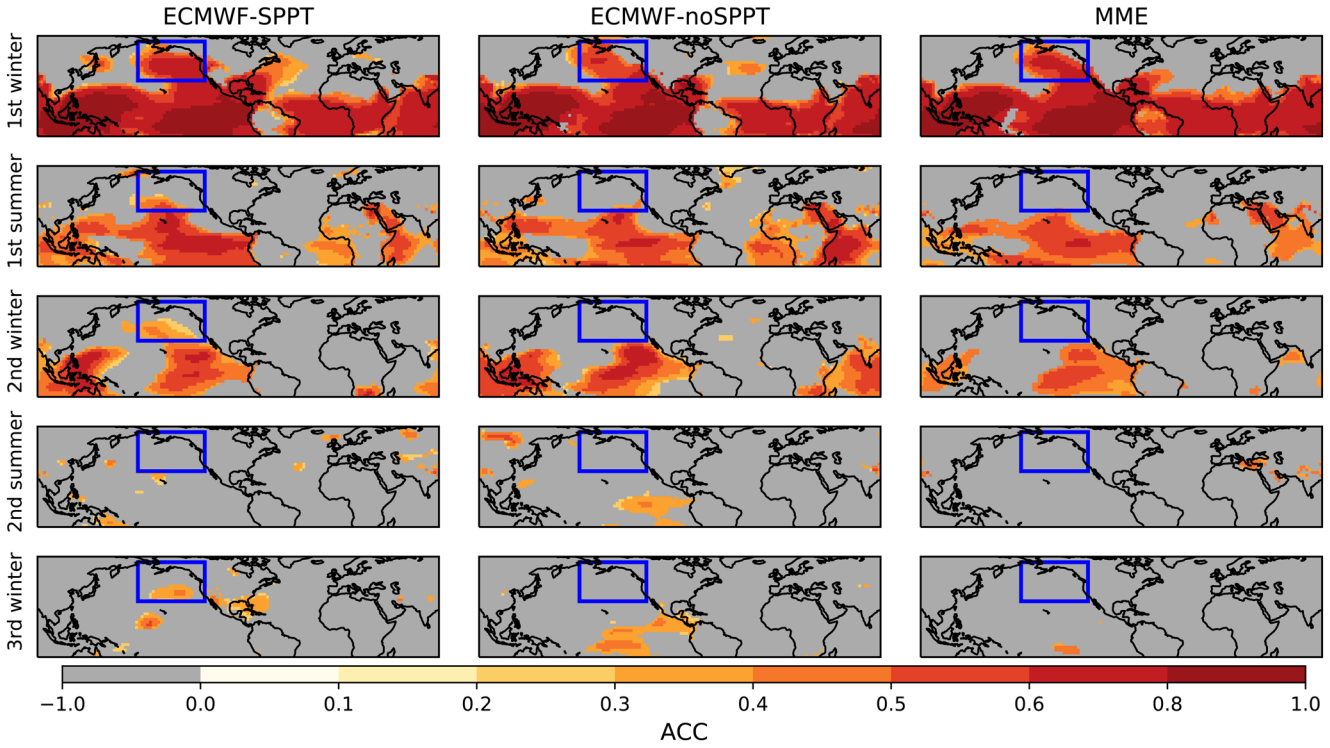

Figure 4. Same as Figure 1 but for SLP. Nongray colors indicate regions with significant positive skill. The blue boxes indicate the regions used to define the North Pacific SLP index (following O'Reilly, 2018). SLP, sea-level pressure.

the *ECMWF-SPPT* ensemble, significant skill returns during the extended second winter. SLP anomalies in this region are strongly influenced by tropical Pacific SST anomalies through an atmospheric teleconnection, so the increased skill in the second winter season is broadly consistent with the larger levels of skill and reliability seen in the *ECMWF-SPPT* ensemble for the NINO3 SST index (Figure 3a).

In addition to the correlation skill, we also analyzed the reliability of the North Pacific SLP index, which is shown in Figure 3d. Neither the *ECMWF-SPPT*, *ECMWF-noSPPT*, or the MME exhibit significant underconfidence or overconfidence for the NPI at any lead time, though this may be limited by the ensemble size. It is clear, however, that in periods around the second winter, all of the constituent models of the MME are overconfident but that the MME itself is reliable, mirroring the behavior seen for the NINO SST indices (Figures 3b and Figures S1b and S2b).

4. Summary

The most pragmatic approach to account for model uncertainty in weather and climate forecasts is the MME. More recently and especially on shorter time scales (up to seasonal), stochastically perturbed ensembles have shown to increase skill and reliability within SMEs. However, such stochastic approaches have only been used to a limited extent in multiannual climate predictions. Here, we have examined the impact of stochastic physics on multiyear time scales, by using targeted model simulations and compared the performance with a MME consisting of the NCAR-DPLE and four CMIP6 dcppA decadal predictions.

Two hindcast sets, performed with and without stochastic physic perturbations (*ECMWF-SPPT* and *ECMWF-noSPPT*), were initialized each November between 1981 and 2014 and run for 28-month forecasts. Results show that SPPT positively impacts reliability for SSTs over the tropical oceans up to about 18 months. The *ECMWF-noSPPT* ensemble is heavily overconfident over the ENSO-relevant NINO3 region, whereas increased reliability is found for the *ECMWF-SPPT* ensemble up to the second winter (DJF; lead time: 14–

16 months) and beyond primarily due to increased ensemble spread. Skill in terms of correlation strength is also higher in the simulation with SPPT up to the second winter.

Furthermore, a comparison of the reliability for SSTs over ENSO regions for six different decadal prediction SMEs has been carried out. Similar to the *ECMWF-noSPPT* IFS ensemble, these decadal models also tend to be overconfident, suggesting that this is a general characteristic of SMEs. In line with previous studies, it is shown that a MME is able to improve the reliability compared to the SMEs, even though these improvements are likely linked to the overconfidence of each single model (Weigel et al., 2008). However, our results suggest that the usage of stochastic physics could be a complementary way to improve the skill and reliability of multiyear predictions.

Assessing the impact of stochastic physics on skill in the simulations used here is difficult over the extratropics due to the small ensemble size of the hindcasts. However, we find evidence that SPPT improves the skill of the large-scale atmospheric circulation over the extratropical North Pacific in the second winter of the forecasts. The improved skill in the extratropics is consistent with being tropically generated and is interpreted here as being a direct result of the improved forecast skill of the tropical Pacific SSTs. Moreover, these forecasts are found to be reliable, in a statistical sense, which increases confidence in the utility of such predictions made with SPPT (Weisheimer & Palmer, 2014). However, further studies based on larger ensembles are needed to fully address the impact of SPPT on multiyear time scales.

Overall, results from this study suggest that stochastic physics represents an effective way to account for model uncertainty in a SME with positive impacts on reliability and also skill on multiyear time scales. Given the low computational costs of these stochastic schemes, it provides a motivation for using them more widely for climate predictions on multiyear time scales, in conjunction with the combination of SMEs into MME.

Data Availability Statement

NCAR-DPLE data can be downloaded from https://www.cesm.ucar.edu/projects/community-projects/DPLE/data-sets.html.

CMIP6 dcppA decadal hindcasts data used in this study can be retrieved via the ESGF (https://esgf-node.llnl.gov/). The authors thank L. Hermanson and D. Smith for providing HadGEM3-GC31-MM simulation data as well as W. Müller and K. Pankatz for providing MPI-ESM1-2-HR simulation data.

Acknowledgments
This study received support from the European Union's Horizon 2020 EUCP project (grant no. GA 776613). Acknowledgment is made for the use of ECMWF's computing and archive facilities in this research, which were provided through the Special Project "Assessing the impact of stochastic physics (SPPT) on sub-decadal timescales." The authors would like to thank two anonymous reviewers for their valuable comments.

References

Andrews, M. B., Ridley, J. K., Wood, R. A., Andrews, T., Blockley, E. W., Booth, B., et al. (2020). Historical simulations with HadGEM3-GC3.1 for CMIP6. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001995. https://doi.org/10.1029/2019MS001995

Batté, L., & Doblas-Reyes, F. J. (2015). Stochastic atmospheric perturbations in the EC-Earth3 global coupled model: Impact of SPPT on seasonal forecast quality. *Climate Dynamics*, *45*(11), 3419–3439. https://doi.org/10.1007/s00382-015-2548-7

Berner, J., Achatz, U., Batté, L., Bengtsson, L., de la Cámara, A., Christensen, H. M., et al. (2017). Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, *98*(3), 565–588. https://doi.org/10.1175/BAMS-D-15-00268.1

Bilbao, R., Wild, S., Ortega, P., Acosta-Navarro, J., Arsouze, T., Bretonnière, P.-A., et al. (2020). Assessment of a full-field initialised decadal climate prediction system with the CMIP6 version of EC-Earth. *Earth System Dynamics Discussions*, *2020*, 1–30. https://doi.org/10.5194/esd-2020-66

Buizza, R., Milleer, M., & Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *125*(560), 2887–2908. https://doi.org/10.1002/qj.49712556006

Cane, M. A. (1991). Forecasting El Niño with a geographical model. *Teleconnections Connecting World-Wide Climate Anomalies*, New York: Cambridge University Press.

Christensen, H. M., Berner, J., Coleman, D. R. B., & Palmer, T. N. (2017). Stochastic parameterization and El Niño-Southern oscillation. *Journal of Climate*, *30*(1), 17–38. https://doi.org/10.1175/JCLI-D-16-0122.1

Corti, S., Weisheimer, A., Palmer, T. N., Doblas-Reyes, F. J., & Magnusson, L. (2012). Reliability of decadal predictions. *Geophysical Research Letters*, *39*, L21712. https://doi.org/10.1029/2012GL053354

Davini, P., von Hardenberg, J., Corti, S., Christensen, H. M., Juricke, S., Subramanian, A., et al. (2017). Climate SPHINX: Evaluating the impact of resolution and stochastic physics parameterisations in the EC-Earth global climate model. *Geoscientific Model Development*, *10*(3), 1383–1402. https://doi.org/10.5194/gmd-10-1383-2017

Doblas-Reyes, F. J., Navarro, J. C. A., Batté, L., Volpi, D., Acosta, M., Bellprat, O., et al. (2018). Using EC-Earth for climate prediction research. *ECMWF Newsletter*, *154*, 35–40. https://doi.org/10.21957/fd9kz3

Doblas-Reyes, F. J., Weisheimer, A., Déqué, M., Keenlyside, N., McVean, M., Murphy, J. M., et al. (2009). Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *135*(643), 1538–1559. https://doi.org/10.1002/qj.464

Haarsma, R., Acosta, M., Bakhshi, R., Bretonnière, P., Caron, L., Castrillo, M., & et al. (2020). HighResMIP versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR – description, model computational performance and basic validation. *Geoscientific Model Development*, *13*, (8), 3507–3527. http://dx.doi.org/10.5194/gmd-13-3507-2020.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*, 1999–2049. https://doi.org/10.1002/qj.3803

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., et al. (2019). SEAS5: The new ECMWF seasonal forecast system. *Geoscientific Model Development*, *12*, 1087–1117. https://doi.org/10.5194/gmd-12-1087-2019

Kataoka, T., Tatebe, H., Koyama, H., Mochizuki, T., Ogochi, K., Naoe, H., et al. (2020). Seasonal to decadal predictions with MIROC6: Description and basic evaluation. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS002035. https://doi.org/10.1029/2019MS002035

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T. K., Balsamo, G., Bechtold, P., et al. (2017). Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, *143*(707), 2315–2339. https://doi.org/10.1002/qj.3094

Lock, S.-J., Lang, S. T. K., Leutbecher, M., Hogan, R. J., & Vitart, F. (2019). Treatment of model uncertainty from radiation by the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme and associated revisions in the ECMWF ensembles. *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 75–89. https://doi.org/10.1002/qj.3570

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019). Developments in the MPI-M earth system model version 1.2 (MPI-ESM1.2) and its response to increasing CO_2. *Journal of Advances in Modeling Earth Systems*, *11*, 998–1038. https://doi.org/10.1029/2018MS001400

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., et al. (2018). A higher-resolution version of the Max Planck institute earth system model (MPI-ESM1.2-HR). *Journal of Advances in Modeling Earth Systems*, *10*, 1383–1413. https://doi.org/10.1029/2017MS001217

O'Reilly, C. H. (2018). Interdecadal variability of the ENSO teleconnection to the wintertime North Pacific. *Climate Dynamics*, *51*, 3333–3350. https://doi.org/10.1007/s00382-018-4081-y

Palmer, T. N. (2001). A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quarterly Journal of the Royal Meteorological Society*, *127*(572), 279–304. https://doi.org/10.1002/qj.49712757202

Palmer, T. N. (2012). Toward the probabilistic Earth-system simulator: A vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *138*(665), 841–861. https://doi.org/10.1002/qj.1923

Palmer, T. N. (2019). Stochastic weather and climate models. *Nature Reviews Physics*, *1*, 463–471. https://doi.org/10.1038/s42254-019-0062-2

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Délécluse, P., et al. (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society*, *85*(6), 853–872. https://doi.org/10.1175/BAMS-85-6-853

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G. J., et al. (2009). Stochastic parametrization and model uncertainty. *ECMWF: Technical Memorandum 598*. https://doi.org/10.21957/ps8gbwbdv

Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., & Smith, L. (2006). Ensemble prediction: A pedagogical perspective. *ECMWF Newsletter*, *106*, 10–17. https://doi.org/10.21957/ab129056ew

Pegion, P., Whitaker, J., Hamill, T., Bates, G., Gehne, M., & Kolczynski, W., Jr. (2016). *Stochastic parameterization development in the NOAA/NCEP global forecast system*. Paper presented at Proceedings of the ECMWF/WWRP Workshop: Model Uncertainty. Reading, UK.

Pohlmann, H., Müller, W. A., Bittner, M., Hettrich, S., Modali, K., Pankatz, K., & Marotzke, J. (2019). Realistic quasi-biennial oscillation variability in historical and decadal hindcast simulations using CMIP6 forcing. *Geophysical Research Letters*, *46*, 14118–14125. https://doi.org/10.1029/2019GL084878

Siegert, S., Bellprat, O., Ménégoz, M., Stephenson, D. B., & Doblas-Reyes, F. J. (2017). Detecting improvements in forecast correlation skill: Statistical testing and power analysis. *Monthly Weather Review*, *145*(2), 437–450. https://doi.org/10.1175/MWR-D-16-0037.1

Smith, D. M., Scaife, A. A., & Kirtman, B. P. (2012). What is the current state of scientific knowledge with regard to seasonal and decadal forecasting? *Environmental Research Letters*, *7*(1), 015602. https://doi.org/10.1088/1748-9326/7/1/015602

Stockdale, T., Doblas-Reyes, F., & Ferranti, L. (2009). EUROSIP: Multi-model seasonal forecasting. *ECMWF Newsletter*, *118*, 10–16. https://doi.org/10.21957/7wc0nybvir

Subramanian, A., Weisheimer, A., Palmer, T., Vitart, F., & Bechtold, P. (2017). Impact of stochastic physics on tropical precipitation in the coupled ECMWF model. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 852–865. https://doi.org/10.1002/qj.2970

Tietsche, S., Balmaseda, M., Zuo, H., Roberts, C., Mayer, M., & Ferranti, L. (2020). The importance of North Atlantic Ocean transports for seasonal forecasts. *Climate Dynamics*, *55*, 1995–2011. https://doi.org/10.1007/s00382-020-05364-6

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C., & Ropelewski, C. (1998). Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *Journal of Geophysical Research*, *103*(C7), 14291–14324. https://doi.org/10.1029/97JC01444

Verfaillie, D., Doblas-Reyes, F. J., Donat, M. G., Pérez-Zanón, N., Solaraju-Murali, B., Torralba, V., & Wild, S. (2020). How reliable are decadal climate predictions of near-surface air temperature? *Journal of Climate*, *34*(2), 697–713. https://doi.org/10.1175/JCLI-D-20-0138.1

Webster, P. J., & Yang, S. (1992). Monsoon and Enso: Selectively interactive systems. *Quarterly Journal of the Royal Meteorological Society*, *118*(507), 877–926. https://doi.org/10.1002/qj.49711850705

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241–260. https://doi.org/10.1002/qj.210

Weisheimer, A., Corti, S., Palmer, T., & Vitart, F. (2014). Addressing model error through atmospheric stochastic physical parameterizations: Impact on the coupled ECMWF seasonal forecasting system. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *372*. https://doi.org/10.1098/rsta.2013.0290

Weisheimer, A., Doblas-Reyes, F. J., Palmer, T. N., Alessandri, A., Arribas, A., Déqué, M., et al. (2009). ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical Research Letters*, *36*, L21711. https://doi.org/10.1029/2009GL040896

Weisheimer, A., & Palmer, T. N. (2014). On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface*, *11*(96), 20131162. https://doi.org/10.1098/rsif.2013.1162

Weisheimer, A., Palmer, T. N., & Doblas-Reyes, F. J. (2011). Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles. *Geophysical Research Letters*, *38*, L16703. https://doi.org/10.1029/2011GL048123

Williams, K. D., Copsey, D., Blockley, E. W., Bodas-Salcedo, A., Calvert, D., Comer, R., et al. (2018). The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations. *Journal of Advances in Modeling Earth Systems*, *10*, 357–380. https://doi.org/10.1002/2017MS001115

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., et al. (2018). Predicting Near-Term Changes in the Earth System: A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model. *Bulletin of the American Meteorological Society*, *99*(9), 1867–1886. https://doi.org/10.1175/BAMS-D-17-0098.1