Diversity of stratospheric error growth across subseasonal prediction systems

Geophysical Research Letters

RESEARCH LETTER
10.1029/2023GL107574

Diversity of Stratospheric Error Growth Across Subseasonal Prediction Systems

R. W. Lee[1] **and A. J. Charlton-Perez**[1]

[1]Department of Meteorology, University of Reading, Reading, UK

Key Points:

- Fitting a detailed statistical model to subseasonal prediction systems reveals some have significant overconfidence in the stratosphere
- Large diversity in stratospheric error growth properties between different systems
- Systems that are overconfident in the stratosphere have an overconfidence bias in the troposphere

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

R. W. Lee,
r.w.lee@reading.ac.uk

Citation:

Lee, R. W., & Charlton-Perez, A. J. (2024). Diversity of stratospheric error growth across subseasonal prediction systems. *Geophysical Research Letters*, *51*, e2023GL107574. https://doi.org/10.1029/2023GL107574

Received 29 NOV 2023
Accepted 13 APR 2024

Author Contributions:

Conceptualization: R. W. Lee, A. J. Charlton-Perez
Data curation: R. W. Lee, A. J. Charlton-Perez
Formal analysis: R. W. Lee, A. J. Charlton-Perez
Investigation: R. W. Lee, A. J. Charlton-Perez
Methodology: R. W. Lee, A. J. Charlton-Perez
Project administration: A. J. Charlton-Perez
Resources: A. J. Charlton-Perez
Software: R. W. Lee, A. J. Charlton-Perez
Supervision: A. J. Charlton-Perez
Validation: R. W. Lee, A. J. Charlton-Perez
Visualization: R. W. Lee, A. J. Charlton-Perez
Writing – original draft: R. W. Lee

Abstract The stratosphere has previously been shown to be a significant source of subseasonal tropospheric predictability. The ability of ensemble prediction systems to appropriately exploit this depends on their ability to reproduce the statistical properties of the real atmosphere. In this study, we investigate predictability properties of the coupled stratosphere-troposphere system in the sub-seasonal to seasonal prediction project hindcasts by fitting a simple, minimal model. We diagnose the signal and noise components of each system in the stratosphere and troposphere and their coupling. We find that while the correlation skill scores are similar in most systems, the signal to noise properties can be substantially different. In the stratosphere, some systems are significantly overconfident, with a quantifiable impact on the tropospheric confidence. We link the method and details of the design of a prediction system to these predictive properties.

Plain Language Summary Subseasonal weather prediction systems make use of ensemble forecasting approaches, in which each forecast is made up of a number of complimentary predictions (ensemble members), that produce more skillful predictions by averaging and allows forecasters to anticipate the uncertainty in any particular forecast. There is a "battle" between the useful signal, hidden in the initial conditions of a simulation run, against the noise that chaotically builds up over the run from inaccuracies in the representation of the initial conditions. We use a simple statistical model with a minimal number of parameters to investigate how well the prediction systems capture both the signal and noise properties of the real atmosphere. We find that while the skill in the troposphere and stratosphere are in line with expectations from the real-world, the ratio between the signal and the noise is too large, particularly in the stratosphere. We find this has an impact on the troposphere below, increasing the signal-to-noise ratio there too—artificially inflating it—giving it an overconfidence which is not wholly statistically or physically justified. We link the method and details of the design of a prediction system to these predictive properties.

1. Introduction

The stratosphere has been shown to be a significant source of tropospheric predictability on subseasonal-to-seasonal (S2S) timescales (see Domeisen et al., 2020a for a review). Numerous other studies have also examined various aspects of average stratospheric-tropospheric coupling, skill, and associated biases (e.g., Domeisen et al., 2020b; Lawrence et al., 2022; Sigmond et al., 2013; Son et al., 2020; Tripathi et al., 2015) with increased surface predictability following weak stratospheric vortex events over the United States, Russia, and the Middle East and some smaller gains in skill following strong vortex events. Ensemble prediction systems (EPSs) with poorly resolved stratospheric processes generally have poorer skill in the troposphere (Domeisen et al., 2020a), although there are no detailed studies which have been able to isolate the quantitative contribution of the stratosphere to tropospheric subseasonal predictive skill. The forthcoming Stratospheric Nudging And Predictable Surface Impacts (SNAPSI) project (Hitchcock et al., 2022) provides a model intercomparison protocol to study this experimentally via control, free and nudged sets of reforecasts for three case study events.

The conclusions of the S2S project show that skill intermittency (the "windows-of-opportunity" concept, Mariotti et al., 2020) is likely to be an important part of the future path to more widespread use of the forecasts. For stratosphere-troposphere coupling, there are a number of factors at play in describing skill intermittency. Several recent studies have suggested that coupling between the stratosphere and troposphere in any given region is strongly dependent on the tropospheric state (e.g., Charlton-Perez et al., 2018; Domeisen et al., 2020c; Kolstad et al., 2022; Lee et al., 2019; Lee et al., 2022; Maycock et al., 2020; Messori et al., 2022). In addition to state dependency, windows of opportunity can also arise serendipitously when the uncertainty in the stratospheric and/or tropospheric forecast is low. While much recent work has focused on examining the mean skill of subseasonal

Writing – review & editing: R. W. Lee, A. J. Charlton-Perez

forecasts in the stratosphere and troposphere, comparatively little attention has been paid to exploring other properties of EPSs.

In this study, we seek to expand this understanding by applying a simple, minimal statistical model to each joint set of forecasts and observations over hindcasts from multiple EPSs in the S2S prediction project archive. The use of a simple, minimal model allows for a deeper investigation of some of these forecasting properties using a clear statistical method. Weigel et al. (2009) and Siegert et al. (2016) are two examples of minimal models which have been used for investigating tropospheric seasonal predictability via signal and noise (and error) terms for the EPSs and corresponding observations. The minimal model of Charlton-Perez et al. (2021), is ideal for our purpose however, because it has the key addition of stratospheric observed and forecast indices, and a term coupling those in the stratosphere with those in the troposphere. This allows for a clear diagnosis of the contribution of the stratosphere to tropospheric forecast skill, as well as other properties such as signal-to-noise ratios (SNRs), and it purposefully excludes state dependent predictability in either the stratosphere or troposphere or in the coupling between the two.

It has been noted by Scaife et al. (2014), Kumar et al. (2014), and Eade et al. (2014) that the SNRs in seasonal predictions are often too low, leading to the "counterintuitive" effect that the EPS is less skillful at predicting members drawn from itself than at predicting the corresponding verification. Statistically however, an anomalous SNR indicates that EPS members are not statistically interchangeable with the verification, and an apparent "paradox" arises only if such an interchangeability is assumed. An anomalous SNR is a consequence of the relative magnitudes of the variance of the observations, the ensemble mean, and the error of the ensemble mean, and should be expected in such circumstances (Bröcker et al., 2023). A Bayesian framework, applied to minimal models, allows the calculation of posterior probabilities for hypotheses, including those related to SNRs (Siegert et al., 2016).

In this study, we set out to compare properties of the coupling and predictability between EPSs in the S2S hindcast data set using a simple, minimal model, fit with Bayesian methods. We will show that there are substantial differences which might be related to how they produce their ensembles and we further show that there is a relation between SNR properties in the stratosphere and troposphere.

2. Simple Minimal Model

Our model, from Charlton-Perez et al. (2021), is as follows:

$$Y_S(t) = \beta_y S(t) + \varepsilon O(t),$$

$$X_{Sk}(t) = \beta_x S(t) + \eta P_k(t) \qquad \text{for } k = 1,\ldots,K,$$

$$Y_T(t) = C_y\, Y_S(t) + \alpha_y T(t) + \lambda Q(t),$$

$$X_{Tk}(t) = C_x X_{Sk}(t) + \alpha_x T(t) + \xi R_k(t) \qquad \text{for } k = 1,\ldots,K.$$

In this model, $Y(t)$ is the observed time series of the parameter of interest for forecasts made at different times, t; $X_k(t)$ are the matching ensemble forecasts. An added subscript S means stratosphere and T means troposphere. $S(t)$, $O(t)$, $P_1(t)$, …, $P_K(t)$, $T(t)$, $Q(t)$, $R_1(t)$, …, $R_K(t)$ are independent standard normal random variables that are also independent over time (i.e., for different t). The model has a predictable "signal" term in both the stratosphere, $S(t)$, and troposphere, $T(t)$, that is identical in the forecasts and observations. These two signal terms are uncorrelated. The noise terms in the stratosphere are $O(t)$, $P_1(t)$, …, $P_K(t)$ and $Q(t)$, $R_1(t)$, …, $R_K(t)$ are the noise terms in the troposphere. The noise terms are uncorrelated with the signal terms and with each other. The two parameters β_y and β_x scale the signal term in the stratosphere, allowing for under- or over-confidence in the forecasts, while α_y and α_x scale the signal term in the troposphere. The ε and η terms similarly scale the noise components in the stratosphere and λ and ξ are the amplitude of the tropospheric noise terms. The correlation between the stratospheric and tropospheric index in the observations is C_y, and for each ensemble member is C_x and is independent of the tropospheric or stratospheric state. See Text S1 in Supporting Information S1 for more details.

The SNRs in the stratosphere and the troposphere are:

$$\mathrm{SNR}_{Sy} = \frac{\beta_y}{\varepsilon},$$

$$\mathrm{SNR}_{Sx} = \frac{\beta_x}{\eta},$$

$$\mathrm{SNR}_{Ty} = \frac{\sqrt{C_y^2\beta_y^2 + \alpha_y^2}}{\sqrt{C_y^2\varepsilon^2 + \lambda^2}},$$

$$\mathrm{SNR}_{Tx} = \frac{\sqrt{C_x^2\beta_x^2 + \alpha_x^2}}{\sqrt{C_x^2\eta^2 + \xi^2}}.$$

This framework can be used to estimate the predictive properties of EPSs by fitting the model parameters from a matched data set of ensemble forecasts and observations using a Bayesian inference approach, as in Siegert et al. (2016). This approach allows the simultaneous estimation of the parameters and quantification of their uncertainty. The approach approximates a fully Bayesian analysis with a Markov chain Monte Carlo (MCMC) integration (Brooks et al., 2011), using the *RStan* package (Stan Development Team, 2019) within *R*. MCMC simulates random draws from an arbitrary probability distribution, such as the posterior distribution:

$$p(\theta, s|x, y) \propto \ell(x, y|\theta, s)\pi(\theta, s)$$

where $\theta = \{\beta_y, \beta_x, \alpha_y, \alpha_x, \varepsilon, \eta, \lambda, \xi, C_y, C_x\}$, the collection of unknown parameters of the signal-plus-noise model; $s = \{S(t), O(t), P_1(t), \ldots, P_K(t), T(t), Q(t), R_1(t), \ldots, R_K(t)\}_{t=1}^{N}$, the unknown values of the latent signal variable; $x, y = \{X_{S1}(t), \ldots, X_{SK}(t), Y_S(t), X_{T1}(t), \ldots, X_{TK}(t), Y_T(t)\}_{t=1}^{N}$, the collection of known forecasts and observations from a hindcast; ℓ is the likelihood function; and π is the prior probability distribution. By using the MCMC sampler, the posterior distributions can be approximated by smoothed histograms and posterior expectations using empirical averages of samples drawn from the posterior distribution. Following the *RStan* default settings, all posterior distributions are sampled using the efficient No-U-Turn Sampler variant of Hamiltonian Monte Carlo (Betancourt, 2018; Hoffman & Gelman, 2014), generated by simulating four parallel Markov chains, each for 2000 iterations, after discarding a spin-up period of 1000 iterations for initialization of the algorithm. All credible intervals are given at the 99% level, throughout.

The prior distribution is a subjective choice in Bayesian analysis, informed by the current state of knowledge. We specified the priors using beta and normal distributions, as detailed in Text S2 in Supporting Information S1. The prior distributions are generally wide and uninformative, and the inference is insensitive to the choice of these prior distributions, with posterior distributions being similar and not changing in any meaningful way. In general, with sufficient data, the influence of the prior disappears, and the Bayesian inference is dominated by ℓ (Gelman & Robert, 2013).

3. Data Sets and Diagnostics

3.1. Ensemble Prediction Systems

The data sets for the indices are taken from hindcasts available from the subseasonal to seasonal project database (Vitart et al., 2017a). Table 1 introduces the main characteristics of the S2S ensemble prediction systems (EPSs) used in this study, including horizontal and vertical resolution, and ensemble strategy. Corresponding "observations" for each EPS are taken from the ERA5 reanalysis (Hersbach et al., 2020). The EPSs were chosen as the most recent set within the S2S project database at the time of commencement of the analysis. While newer versions of most EPSs now exist, many of the same elements remain including various initialization and perturbation strategies as well as a variety of resolutions. There are substantial differences in the construction and configuration between the EPSs (Merryfield et al., 2020; Takaya, 2019), including methods to represent the uncertainty in initial conditions (initialization strategy—lagged ensemble with different initial times or burst ensemble with a common initial time) and model physics, resolution (horizontal and vertical), ensemble size,

Geophysical Research Letters
10.1029/2023GL107574

Table 1
Details of the Ensemble Prediction Systems Considered

Symbol	Center	Version	Horizontal resolution	Levels (model top)	Hindcast					Initialization strategy	Perturbations strategy
					Period	Frequency	Length [days]	Size	Merged size		
●	ECMWF (ecmf)	Cy45R1	Tco639 (~16 km) up to day 15; Tco319 (~32 km) after day 15	91 (0.01 hPa)	1999–2019	2/week	46	11		Burst mode (singular vectors; ensemble of data assimilation)	Model physics: SPPT and SKEB
■	UKMO (egrr)	GloSea5 GC2	N216 0.83° × 0.56° (~60 km in mid-latitudes)	85 (85 km)	1993–2015	4/month	60	7		Lagged mode (4/day in forecast)	Model physics: SKEB
◆	NCEP (kwbc)	CFSv2	T126 (~100 km)	64 (0.02 hPa)	1999–2010	6 hourly	44	1	12	Lagged mode	-
×	CNRM (lfpw)	CNRM-CM 6.0	TL255 (~80 km)	91 (0.01 hPa)	1993–2014	4/month	61	15		Burst mode (but no initial spread)	Model dynamics: (Batté & Déqué, 2012)
A	BoM (ammc)	POAMA P24	T47 (~250 km)	17 (10 hPa)	1981–2013	6/month	~270	33		Burst mode (coupled bred vectors)	Model physics: 3 model versions (2 shallow convection, 1 flux adjustment); Model dynamics
B	CMA (babj)	BCC-CPS-S2Sv1	T106 (about 110 km)	40 (0.5 hPa)	1994–2014	Daily	60	4	12	Lagged mode	-
C	KMA (rksl)	Glosea5 GC2	N216 0.83° × 0.56° (~60 km in mid-latitudes)	L85 (85 km)	1991–2010	4/month	≤240	3		Lagged mode (4/day in forecast)	Model physics: SKEB2
D	HMCR (rums)	RUMS	1.125° × 1.40625°	28 (5 hPa)	1991–2015	Weekly	61	10		Burst mode (breeding vectors)	–
E	JMA (rjtd)	GEPS1701	TL479 (~40 km) up to day 18; TL319 (~55 km) after day 18.	100 (0.01 hPa)	1981–2012	3/month	34	5		All (LETKF; singular vectors; lagged averaging forecasts; SST perturbations)	Model physics: SPPT
F	ECCC (cwao)	GEPS5	Yin-Yang grid at 0.35° (~39 km)	45 (0.1 hPa)	1998–2017	Weekly	32	4		Burst mode (Ensemble Kalman Filter)	Model physics: multi-parameterization, stochastic perturbations, SKEB

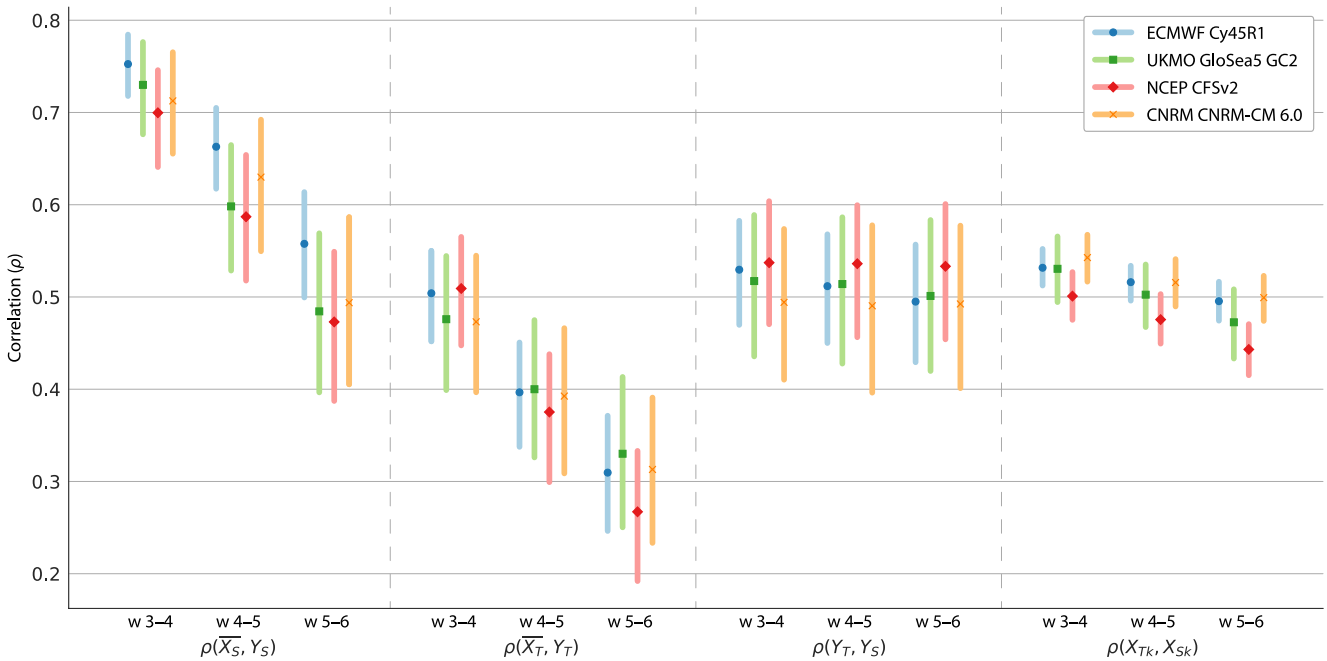

Figure 1. Correlations, as measured by the Pearson's correlation, ρ, of subseasonal forecasts. Showing the ensemble mean correlations in the stratosphere, $\rho(\overline{X_S}, Y_S)$, and troposphere, $\rho(\overline{X_T}, Y_T)$; and the stratosphere-troposphere correlations in the observations, $\rho(Y_T, Y_S)$, and forecasts, $\rho(X_T, X_S)$. Each set of correlations are shown for weeks (w) 3–4, 4–5, and 5–6. Correlations are calculated by first taking the mean value of the index over these days. $\rho(X_T, X_S)$ are the ensemble mean of the correlation in each individual ensemble member. Credible intervals are 99%.

ensemble perturbation strategy (in-run adjustments), and hindcast period. These differences impact forecast quality and our ability to evaluate the performance of the hindcast (Merryfield et al., 2020).

3.2. Stratospheric and Tropospheric Indices

The minimal model uses standardized stratospheric and tropospheric climate indices for forecast and observations. As a representative stratospheric index we use the stratospheric Northern Annular Mode (NAM) in the lower stratosphere at 100 hPa, where forecast skill is high (Son et al., 2020), derived using the zonal mean principal component method of Baldwin and Thompson (2009). As a representative tropospheric index we use the North Atlantic Oscillation (NAO), here defined as the mean sea-level pressure difference between a 2.5° × 2.5° grid box centered at 65°N, 20°W (Iceland), 37.5°N, 25°W (Azores). The lead time dependent bias is removed, prior to analysis. With biweekly temporal grouping used throughout, "weeks 3–4" here represents days 14–27, "weeks 4–5" represents days 21–34, and "weeks 5–6" represents days 28–41 into each forecast. All hindcasts in the database initialized between November–February for the particular model version in question are considered. No attempt is made to standardize the period over which the forecasts are made. For the NCEP model, a lagged ensemble is created by combining forecasts initialized over three consecutive days, producing an ensemble of similar size to the other hindcasts analyzed.

The following sections use the minimal model to assess and contrast the S2S EPSs, focusing on skill, signal and noise components of each system in the stratosphere and troposphere and their coupling.

4. Correlation Skill

Before considering the signal to noise properties of the systems, we first analyses their ability to produce skillful forecasts. The forecast skill in the stratosphere and troposphere is largely in line with previous studies (e.g., Domeisen et al., 2020b) and skill does not vary to any significant degree between the modeling systems tested. The forecast skill (Figure 1), as measured by the Pearson's correlation derived from the minimal model (Text S1 in

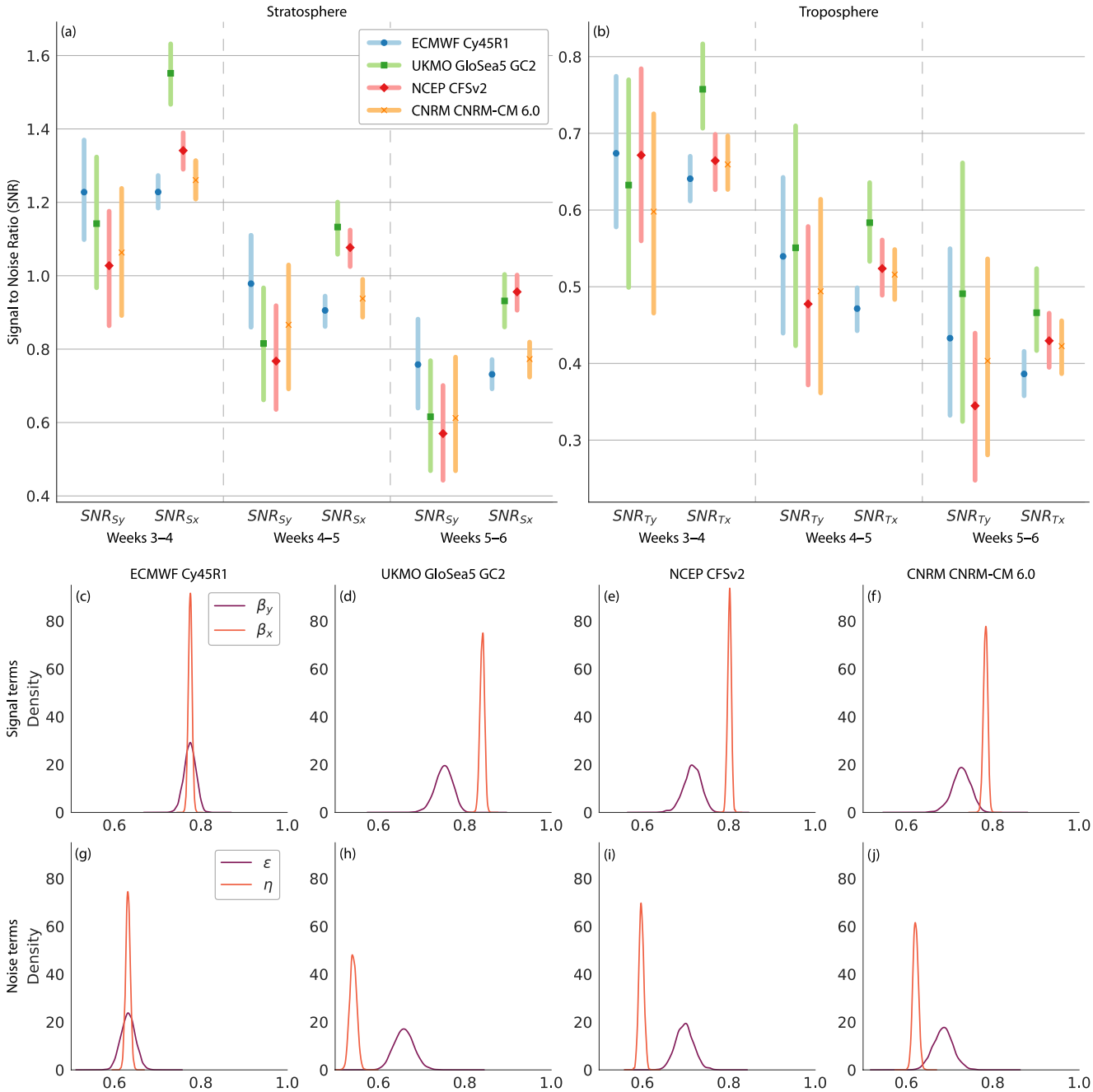

Figure 2. (a–b): Signal-to-noise ratios, *SNR*, in the (a) stratosphere ($_S$) and (b) troposphere ($_T$) in the observations ($_y$) and model forecasts ($_x$) for weeks 3–4, 4–5 and 5–6. Credible intervals are 99%. (c–j): Stratospheric forecast parameters for: the signal terms, β_y and β_x for the observations and forecasts respectively (middle row); and the noise terms, ε and η (lower row), for the observations and forecasts respectively at 3–4 weeks lead time. EPSs: (c) and (b) ECMWF, (d) and (h) UKMO, (e) and (i) NCEP, (f) and (j) CNRM.

Supporting Information S1), between EPS ensemble means and observations decreases with lead time over the 14-day averaged weeks 3–4, 4–5, and 5–6, as expected, in both the stratosphere ($\rho(\overline{X_S}, Y_S)$) and troposphere ($\rho(\overline{X_T}, Y_T)$). In addition, correlations are higher in the stratosphere than troposphere, and EPS mean correlations are all relatively similar to each other.

In terms of stratosphere-troposphere coupling (Figure 1), The posterior mean of the correlation between the stratospheric and tropospheric state (C_x and C_y) is not distinguishably different between the observations and forecast system for any of the four systems considered. Correlations in the observations ($\rho(Y_T, Y_S)$) remain constant with lead time, as would be expected, with variations between lead times due to sampling, and variations between models due to sampling and natural variability (considering different hindcast periods). Stratosphere-troposphere coupling in the EPSs ($\rho(X_{Tk}, X_{Sk})$) is consistent with the correlation in the observations in all EPSs except NCEP at weeks 4–5 and 5–6 where it is significantly lower ($Pr(\rho(X_T, X_S) > \rho(Y_T, Y_S))$ is 0.019 and 0.003, respectively; Table S1 in Supporting Information S1), nonetheless suggesting that there is little difference in the strength of stratosphere-troposphere coupling between the models. This conclusion is important because it implies that any deficiency in the EPSs' tropospheric predictions is not linked to a lack of coupling between the stratosphere and the troposphere.

5. Signal and Noise

By comparing the posterior estimates of signal-to-noise ratio (SNR) in the observations and forecasting systems it is possible to determine if forecasts produced by the forecasting system are over or under confident. While on seasonal timescales some EPSs are underconfident (Baker et al., 2018; Eade et al., 2014; Scaife et al., 2014; Siegert et al., 2016), here we will show that on the subseasonal timescale some forecasting systems are over-confident (and therefore likely under dispersive) in their stratospheres (and to a much lesser extent also in their tropospheres), stemming from too much signal and too little noise.

Stratospheric SNRs (Figure 2a) in the ECMWF EPS are consistent ($0.05 > \Pr(\mathrm{SNR}_{Sx} > \mathrm{SNR}_{Sy}) < 0.95$) with their corresponding observations across all lead times. Estimates of the SNR in the observations are similar for all four periods. However SNRs in the UKMO, NCEP, and sometimes CNRM EPSs are significantly larger than observed —stratospheric forecasts in these systems are overconfident. There is very high confidence (Table S2 in Supporting Information S1) that this is the case for all weeks in the UKMO and NCEP EPSs ($\Pr(\mathrm{SNR}_{Sx} > \mathrm{SNR}_{Sy}) = 1$), and for weeks 3–4 and 5–6 in the CNRM EPS ($\Pr(\mathrm{SNR}_{Sx} > \mathrm{SNR}_{Sy}) = 0.999$ and 0.994, respectively).

The stratospheric signal (β_y, β_x) and noise (ε, η) parameters from the minimal model are shown next for weeks 3–4 lead time (Figures 2c–2j), to investigate the origins of the anomalously large SNRs. In common with the correlations and SNRs there are posterior distribution differences, with the observations featuring a larger variance due to less data relative to the modeling systems each of which have more ensemble members. As implied from the SNRs, the ECMWF system features very similar central tendency of their signal and noise distributions with their corresponding observations. The CNRM system has a larger signal and smaller noise relative to their corresponding observations, with just the far tails of the two distributions overlapping in each case ($Pr(\beta_x > \beta_y) = 0.999$). UKMO and NCEP EPSs feature entirely separated distributions compared to their corresponding observations for both signal (larger) and noise (smaller), again indicating the initializations may be under-dispersed in the stratosphere. This same pattern of relative differences for the different systems and their corresponding observations qualitatively remains similar over weeks 4–5 and 5–6 (not shown), with the main absolute differences being a shift toward a smaller signal and increased noise in both the systems and their corresponding observations.

The tropospheric SNR (Figure 2b) posterior distributions from all EPSs overlap with their corresponding observations at all lead times, to varying degrees. The UKMO at week 3–4 ($\Pr(\mathrm{SNR}_{Tx} > \mathrm{SNR}_{Ty}) = 0.982$) and NCEP at week 5–6 ($\Pr(\mathrm{SNR}_{Tx} > \mathrm{SNR}_{Ty}) = 0.978$) are the least consistent (see Table S2 in Supporting Information S1 for all probabilities).

Overall, the EPSs which run in lagged ensemble mode (Table 1), UKMO and NCEP, feature significantly larger SNRs, via both increased signals and reduced noise, relative to the EPSs which run in burst mode with stochastic spread of initial conditions, namely ECMWF. While the CNRM EPS is also run in burst mode, it relies only on in-run perturbations to model dynamics, choosing not to perturb initial conditions. Therefore, by weeks 3–4 there are also some problems with a high signal and reduced noise in the stratosphere, but not nearly as severe as the lagged ensembles. It therefore appears that the method of ensemble perturbation generation might in some EPSs yield an overconfident, under-dispersed stratosphere on subseasonal timescale, with too much signal and too little noise, stemming from insufficient spread in initial conditions to represent the true uncertainty. Such lagged ensemble methods for initial condition generation may have been sufficiently well assessed for their SNR in the troposphere

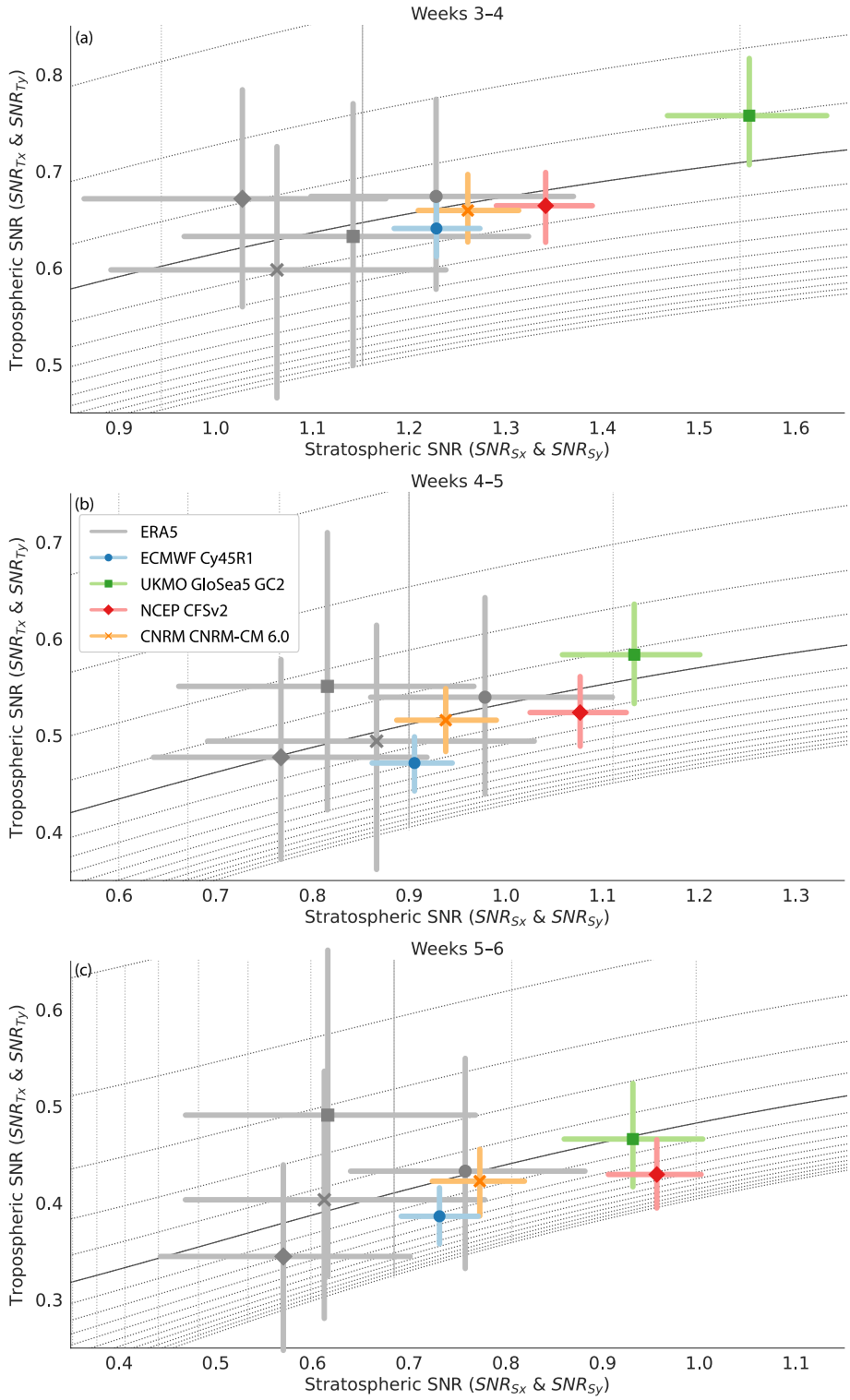

Figure 3. Signal-to-noise ratios, *SNR*, in the stratosphere ($_S$) versus the troposphere ($_T$) in the observations (gray; $_y$) and forecasts (colors; $_x$) at (a) 3–4, (b) 4–5, and (c) 5–6 weeks lead time. Each sample of the observations is matched to its respective EPS forecast by shape. Credible intervals are 99%. Curved dark gray background lines represents the fractional change in the size of the stratospheric signal between the EPS and observations (an EPS-observation pair parallel to a line means differences are entirely determined by the stratospheric SNR difference). Vertical light gray background lines represent the fractional change in the size of the tropospheric signal between the EPS and observations (an EPS-observation pair parallel to a line means differences are entirely determined by the tropospheric SNR difference) (see Text S3 and Figure S1 in Supporting Information S1 for a more detailed explanation).

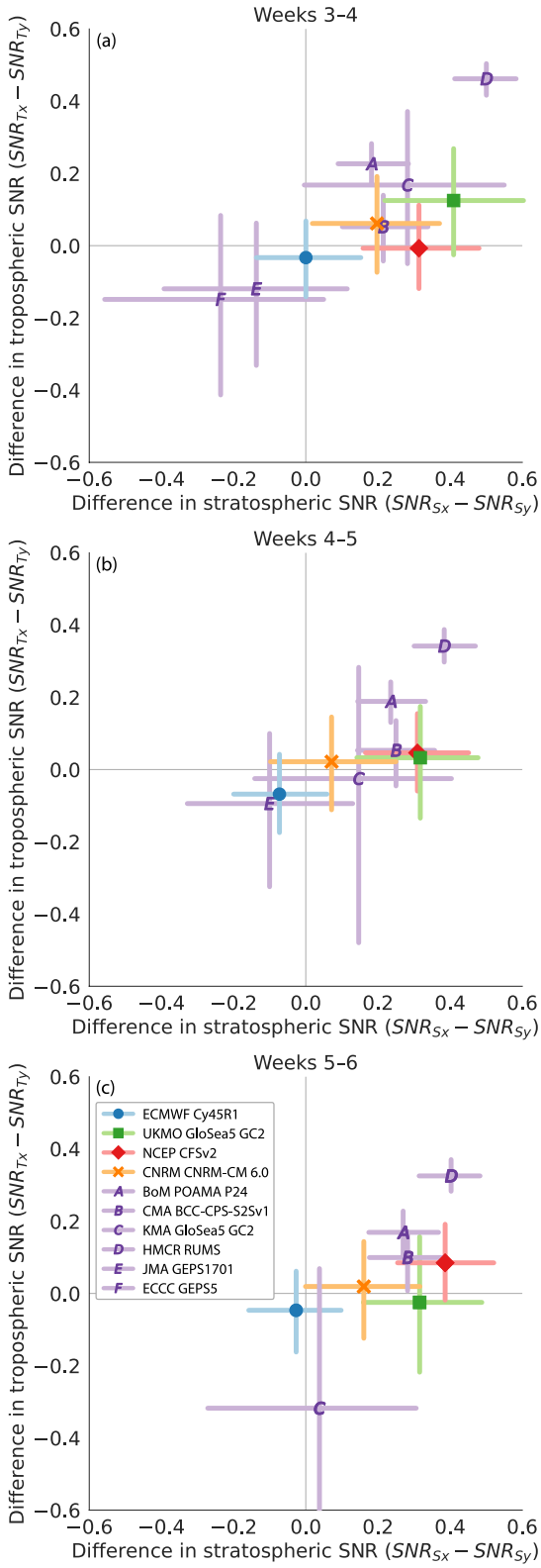

Figure 4. Differences in signal-to-noise ratios, *SNR*, in the stratosphere ($_S$) versus the troposphere ($_T$) between the model forecasts ($_x$) and their respective observations ($_y$) at (a) 3–4, (b) 4–5, and (c) 5–6 weeks lead time. Credible intervals are 99%.

where growth timescales are faster, but not in the stratosphere where they are slower. However, there are many other differences between the systems.

6. Discussion and Conclusions

Might an overly confident stratosphere bias the signal-to-noise in the troposphere? To analyze this, the SNR is shown for the stratosphere against the troposphere for weeks 3–4, 4–5 and 5–6 (Figure 3) with the addition of theoretical minimal model curves of fixed stratospheric SNR (Text S3 in Supporting Information S1). Displayed in this way, it is again clear that in the UKMO system the stratospheric SNR is outside of the corresponding observations at weeks 3–4. Furthermore, taking the example of weeks 3–4 (Figure 3a), if we were to correct the stratospheric SNR bias (say via a relevant EPS upgrade) and the UKMO system would shift left, parallel to the background guideline curves, to give it the same (*x*-axis, SNR_S) value as the corresponding observations, we see the UKMO system now only weakly positively biased for the tropospheric SNR (*y*-axis, SNR_T). This suggests that the stratosphere is the leading cause of the tropospheric SNR bias. The NCEP system is also outside of the corresponding observations (Figure 3c).

To investigate if other S2S EPSs follow the same overall pattern of an association between stratospheric and tropospheric SNR, the differences are computed for the other S2S systems in the database in addition to the four already considered and their corresponding observations for up to six extra EPSs (Figure 4, Table 1). By adding these additional EPSs, some of which are highly overconfident (e.g., HMCR), while others are underconfident and over-dispersed (e.g., ECCC), they still maintain the same diagonal line grouping across all EPSs, and across all lead times. This confirms that EPSs that feature an overly confident stratosphere have a quantifiable impact on the troposphere on the extended range/subseasonal timescale. This study uses an ensemble of opportunity where the EPSs differ in many ways: model dynamics, ensemble strategy, resolution, ensemble size (Table 1). We do note, however, that many outlying EPSs tend to feature lagged initialization ensemble strategies and/or under resolved stratospheres (low model top, low number of vertical levels) which would tend to curtail stratospheric variability and may contribute to overconfidence. EPSs with few ensemble members manifest themselves in Figure 4 with a large confidence interval. There may also be other attributes of EPS design leading to such SNR biases; investigation into the causes warrants further work.

To make use of predictability from all aspects of the system, including the stratosphere, EPS design should therefore, in addition to having sufficient vertical and horizontal resolution to simulate important processes of subseasonal predictability, also carefully consider ensemble perturbation techniques to generate sufficient spread to represent the true uncertainty where the growth of noise/errors is slower. At longer (>6 weeks) lead times, this may become less important as tropospheric noise (and in-run perturbations, where applied) eventually reduce the overconfidence and particularly its impact on the tropospheric SNR. This also implies that the low SNR on seasonal-to-decadal timescales (e.g., Eade et al., 2014) may be somewhat masked on the seasonal end of this range by those overconfident stratospheres in some EPSs, and therefore emerge later in the forecast lead time than they would otherwise if the EPSs were not under-dispersive at initialization and shorter lead times.

Data Availability Statement

This work is based on S2S data. S2S is a joint initiative of the World Weather Research Programme (WWRP) and the World Climate Research Programme (WCRP). The original S2S database is hosted at ECMWF (Vitart et al., 2017b) as an extension of the TIGGE database. ERA5 hourly data on pressure and surface levels used in the study was obtained from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS) at Hersbach et al. (2023a, 2023b). The RStan software used in the study is available at Stan Development Team (2019).

Acknowledgments
We sincerely thank the two anonymous reviewers for their constructive comments which have helped to improve the manuscript. This research was supported by the University of Reading.

References

Baker, L. H., Shaffrey, L. C., Sutton, R. T., Weisheimer, A., & Scaife, A. A. (2018). An intercomparison of skill and overconfidence/under-confidence of the wintertime North Atlantic Oscillation in multimodel seasonal forecasts. *Geophysical Research Letters*, *45*(15), 7808–7817. https://doi.org/10.1029/2018gl078838

Baldwin, M. P., & Thompson, D. W. J. (2009). A critical comparison of stratosphere–troposphere coupling indices. *Quarterly Journal of the Royal Meteorological Society*, *135*(644), 1661–1672. https://doi.org/10.1002/qj.479

Batté, L., & Déqué, M. (2012). A stochastic method for improving seasonal predictions. *Geophysical Research Letters*, *39*(9), L09707. https://doi.org/10.1029/2012GL051406

Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo (Version 2). *arXiv, 1701.02434*. https://doi.org/10.48550/ARXIV.1701.02434

Bröcker, J., Charlton–Perez, A. J., & Weisheimer, A. (2023). A statistical perspective on the signal-to-noise paradox. *Quarterly Journal of the Royal Meteorological Society*, *149*(752), 911–923. https://doi.org/10.1002/qj.4440

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (Eds.) (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC. https://doi.org/10.1201/b10905

Charlton-Perez, A. J., Bröcker, J., Karpechko, A. Y., Lee, S. H., Sigmond, M., & Simpson, I. R. (2021). A minimal model to diagnose the contribution of the stratosphere to tropospheric forecast skill. *Journal of Geophysical Research: Atmospheres*, *126*(24), e2021JD035504. https://doi.org/10.1029/2021jd035504

Charlton-Perez, A. J., Ferranti, L., & Lee, R. W. (2018). The influence of the stratospheric state on North Atlantic weather regimes. *Quarterly Journal of the Royal Meteorological Society*, *144*(713), 1140–1151. https://doi.org/10.1002/qj.3280

Domeisen, D. I. V., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., et al. (2020a). The role of the stratosphere in subseasonal to seasonal prediction: 2. Predictability arising from stratosphere-troposphere coupling. *Journal of Geophysical Research: Atmospheres*, *125*, e2019JD030923. https://doi.org/10.1029/2019jd030923

Domeisen, D. I. V., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., et al. (2020b). The role of the stratosphere in subseasonal to seasonal prediction: 1. Predictability of the stratosphere. *Journal of Geophysical Research: Atmospheres*, *125*(2), e2019JD030920. https://doi.org/10.1029/2019jd030920

Domeisen, D. I. V., Grams, C. M., & Papritz, L. (2020c). The role of North Atlantic–European weather regimes in the surface impact of sudden stratospheric warming events. *Weather and Climate Dynamics*, *1*(2), 373–388. https://doi.org/10.5194/wcd-1-373-2020

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*(15), 5620–5628. https://doi.org/10.1002/2014gl061146

Gelman, A., & Robert, C. P. (2013). "Not only defended but also applied": The perceived absurdity of bayesian inference. *The American Statistician*, *67*(1), 1–5. https://doi.org/10.1080/00031305.2013.760987

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., et al. (2023a). ERA5 hourly data on pressure levels from 1940 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* [Dataset]. https://doi.org/10.24381/cds.bd0915c6

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., et al. (2023b). ERA5 hourly data on single levels from 1940 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* [Dataset]. https://doi.org/10.24381/cds.adbb2d47

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*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hitchcock, P., Butler, A., Charlton-Perez, A., Garfinkel, C. I., Stockdale, T., Anstey, J., et al. (2022). Stratospheric nudging and predictable surface impacts (SNAPSI): A protocol for investigating the role of stratospheric polar vortex disturbances in subseasonal to seasonal forecasts. *Geoscientific Model Development*, *15*(13), 5073–5092. https://doi.org/10.5194/gmd-15-5073-2022

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(47), 1593–1623. Retrieved from https://jmlr.org/papers/v15/hoffman14a.html

Kolstad, E. W., Lee, S. H., Butler, A. H., Domeisen, D. I. V., & Wulff, C. O. (2022). Diverse surface signatures of stratospheric polar vortex anomalies. *Journal of Geophysical Research: Atmospheres*, *127*(20), e2022JD037422. https://doi.org/10.1029/2022jd037422

Kumar, A., Peng, P., & Chen, M. (2014). Is there a relationship between potential and actual skill? *Monthly Weather Review*, *142*(6), 2220–2227. https://doi.org/10.1175/mwr-d-13-00287.1

Lawrence, Z. D., Abalos, M., Ayarzagüena, B., Barriopedro, D., Butler, A. H., Calvo, N., et al. (2022). Quantifying stratospheric biases and identifying their potential sources in subseasonal forecast systems. *Weather and Climate Dynamics*, *3*, 977–1001. https://doi.org/10.5194/wcd-3-977-2022

Lee, S. H., Charlton-Perez, A. J., Woolnough, S. J., & Furtado, J. C. (2022). How do stratospheric perturbations influence North American weather regime predictions? *Journal of Climate*, *35*(18), 5915–5932. https://doi.org/10.1175/jcli-d-21-0413.1

Lee, S. H., Furtado, J. C., & Charlton-Perez, A. J. (2019). Wintertime North American weather regimes and the Arctic stratospheric polar vortex. *Geophysical Research Letters*, *46*(24), 14892–14900. https://doi.org/10.1029/2019gl085592

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., Collins, D. C., et al. (2020). Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bulletin of the American Meteorological Society*, *101*(5), E608–E625. https://doi.org/10.1175/bams-d-18-0326.1

Maycock, A. C., Masukwedza, G. I. T., Hitchcock, P., & Simpson, I. R. (2020). A regime perspective on the North Atlantic eddy-driven jet response to sudden stratospheric warmings. *Journal of Climate*, *33*(9), 3901–3917. https://doi.org/10.1175/jcli-d-19-0702.1

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A. S., et al. (2020). Current and emerging developments in sub-seasonal to decadal prediction. *Bulletin of the American Meteorological Society*, *101*(6), E869–E896. https://doi.org/10.1175/bams-d-19-0037.1

Messori, G., Kretschmer, M., Lee, S. H., & Wendt, V. (2022). Stratospheric downward wave reflection events modulate North American weather regimes and cold spells. *Weather and Climate Dynamics*, *3*(4), 1215–1236. https://doi.org/10.5194/wcd-3-1215-2022

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., et al. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, *41*(7), 2514–2519. https://doi.org/10.1002/2014gl059637

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*(3), 995–1012. https://doi.org/10.1175/jcli-d-15-0196.1

Sigmond, M., Scinocca, J. F., Kharin, V. V., & Shepherd, T. G. (2013). Enhanced seasonal forecast skill following stratospheric sudden warmings. *Nature Geoscience*, *6*(2), 98–102. https://doi.org/10.1038/ngeo1698

Son, S., Kim, H., Song, K., Kim, S., Martineau, P., Hyun, Y., & Kim, Y. (2020). Extratropical prediction skill of the subseasonal-to-seasonal (S2S) prediction models. *Journal of Geophysical Research: Atmospheres*, *125*(4), e2019JD031273. https://doi.org/10.1029/2019jd031273

Stan Development Team. (2019). RStan: The R interface to stan (2019, October 19. Version 2.21.0) [Software]. Retrieved from https://mc-stan.org/

Takaya, Y. (2019). Forecast system design, configuration, and complexity. In A. W. Robertson & F. Vitart (Eds.), *Sub-seasonal to seasonal prediction* (pp. 245–259). Elsevier. https://doi.org/10.1016/b978-0-12-811714-9.00012-7

Tripathi, O. P., Charlton-Perez, A., Sigmond, M., & Vitart, F. (2015). Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions. *Environmental Research Letters*, *10*, 104007. https://doi.org/10.1088/1748-9326/10/10/104007

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al. (2017a). The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/bams-d-16-0017.1

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al. (2017b). The subseasonal to seasonal (S2S) prediction project database [Dataset]. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/bams-d-16-0017.1

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2009). Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Monthly Weather Review*, *137*(4), 1460–1479. https://doi.org/10.1175/2008mwr2773.1